

Not everything is in the genes

Last week's reports of a possible genetic basis to an individual's ability to handle complex social situations highlights the importance of developing frameworks to ensure that genetic information is used responsibly.

“Why boys will be boys.” “Girls may inherit intuition gene from father.” The newspaper headlines that accompanied last week's report about research by David Skuse and his colleagues at the Institute of Child Health in London, suggesting a strong, sex-dependent genetic component in reactions to certain types of social situations, had a certain predictability (see *Nature* 387, 705; 1997). The explicit message that these headlines, and the articles that accompanied them, conveyed was dramatic, describing how the research appears to have produced a clear example of the way in which modern genetics may provide us with useful tools for understanding social behaviour.

Skuse studied a group of 80 girls with Turner's syndrome, characterized by possessing only one X chromosome (females usually have a pair). Some had inherited their single X chromosome from their father and some from their mother, and the study found that the parental origin of the X chromosome was closely correlated with an individual girl's ability to cope with social situations. Similar differences were found between boys — who have a single X chromosome inherited from their mother — and girls. Skuse's conclusion, widely quoted in the press reports, was that his results show a clear genetic basis for the difference in social skills between women and men.

These results are scientifically stimulating. But it is important to remain wary of over-interpretation. For there is an inevitable temptation to use them to justify a form of genetic determinism which — despite our awareness of the history of eugenics movements and other misuses of genetics in the past — remains a powerful force in modern society. Reporting the hunt for, or discovery of, a ‘gene for’ homosexuality, violence, alcoholism or mental illness has become a regular feature of newspaper science. Each time it happens, there are scientists quick to point out how the characteristics described are frequently the result of a complex interaction between genetic susceptibility and environmental stimulus. But there are a few others, more provocatively

quotable, who are equally quick to use such ‘discoveries’ to support claims of the extent to which the straitjacket of nature remains impervious to the coaxing of nurture.

The very successes of genetics research can, often unwittingly, appear to strengthen the latter arguments, and thus the social policies that appear to flow logically from them. The myth of biological utopia reached its nadir in the eugenic policies of Nazi Germany. Few believe the crude myth any longer. But, in more sophisticated forms, it remains a powerful influence on certain trends of modern political and social thought. Of course, where there is solid and credible evidence of a biological basis to certain aspects of social behaviour, it would be foolish to ignore the implications. Knowledge of such biological roots, for example, could play an important role in helping individuals to cope with difficult social situations more effectively.

The danger, however, is that the biological or genetic explanation becomes the dominant measure of ‘normality’, and that other more subtle influences on behaviour are denied their legitimate role. Perhaps this danger is most acute in the study of mental illness, where sufferers from diseases found to have a strong genetic component risk being labelled as socially ‘subnormal’. But similar threats exist for anyone who comes to be recognized as ‘genetically disabled’ — with the implication that society has little chance of improving their plight.

Ironically, the new insights into social behaviour provided by modern genetics make it more, not less, important that we learn to imbue our use of these insights with a basic commitment to human rights. The proposed universal declaration on the human genome being drawn up by the United Nations Educational, Scientific and Cultural Organization is a timely step towards modernizing legislation in this field, both nationally and internationally. Modern genetics has provided one of the most vivid illustrations of Francis Bacon's maxim that “knowledge is power”. We share a common responsibility to use that knowledge for the benefit of all. □

A lesson on preparedness

Guidelines for handling allegations of scientific fraud are essential to preserve the reputation of science.

Germany has been caught off guard. The emergence of a scientific fraud case of substantial proportions (see page 750) has exposed the fact that few research institutions have formal guidelines for dealing with allegations of scientific misconduct. As a result, Germany's scientific community has had to respond in an *ad hoc* way to the new case, involving two researchers, now at different universities, who stand accused of fabricating results when they worked together at a research institute in Berlin.

So far, the investigation into the allegations appears to have been carried out in an exemplary fashion. Committees of inquiry were rapidly set up at all three institutions, and their conclusions were put to a common committee of experts last week. The deans of the two universities and the director of the Berlin research centre have discussed their investigations openly. Both universities have been in close contact with their respective regional ministries of education, which are responsible for hiring or firing university faculty members.

Perhaps it has needed a case like this one to rouse German academics from their traditional complacency in regarding fraud as something to be handled discreetly by the research departments concerned. Indeed, many leading academics have argued in the past that scientific fraud could never become a significant problem in Germany, where the struggle for research grants is much less vicious than in the United States. The few lonely voices which have been calling for guidelines to be prepared — in case of need — are now finding keener ears.

The outcome of this unprecedented case will help to define what can be done within the constraints of Germany's legal system. A US-style central office of scientific integrity is not a solution in a country whose research funding organizations prize their independence from the state. But the heads of Germany's research organizations and its rectors' conference should at least agree on a common set of guidelines adaptable to individual needs. Such guidelines would protect not only the integrity of German science, but also its reputation with the taxpayer. □



NIH backers take Clinton to task over impact of budget pact

[WASHINGTON] The potential ramifications for biomedical research of the balanced budget agreement recently struck by President Bill Clinton and congressional leaders have prompted a groundswell of concern from supporters of the US National Institutes of Health (NIH).

In a document prepared at the end of last month, the Office of Management and Budget (OMB) estimates that, under the terms of the budget agreement, the NIH would be held to a 1.2 per cent increase in 1998 over its 1997 funding. That would be needed to allow funding for other, non-biomedical, spending explicitly protected in the budget deal.

Such an increase is far below the projected 3.1 per cent rate of biomedical inflation calculated by the NIH, based on the historical relationship between the general rate of inflation and inflation for biomedical goods and services. It also represents less than half of the 2.6 per cent increase contained in Clinton's formal budget request for 1998, released in February.

At the heart of the resulting criticism of these figures is the fact that, while Clinton has publicly and frequently applauded biomedical research, he did not make it one of the explicitly "protected areas" of domestic spending in the agreement he struck with Republican leaders of Congress last month.

This agreement would eliminate the US budget deficit by 2002, while protecting five areas of domestic spending. These include education and the environment and, within these, specific programmes covering subjects from national parks to immigrant education.

But the result is that the rest of the limited funds for domestic spending is tightly constrained — including that for NIH. "Clearly biomedical research was not treated as a priority and is very exposed because so many other things were treated as priorities by the president," said one biomedical lobbyist.

The OMB document estimates that, under the budget agreement, the NIH budget would need to be held to less than \$12.9 billion, rather than growing from \$12.7 billion in 1997 to \$13.1 billion in 1998 as Clinton requested in February.

Shortly after this estimate was sent to Congress, OMB officials distanced themselves from the figure, calling it a "mechanical" calculation. "It was not reviewed by any administration policy officials, and it should not be construed in any way to reflect the administration's position," Barry Anderson, OMB's assistant director for budget, wrote to the House of Representatives Appropriations Committee.

But the critics soon struck. John Porter (Republican, Illinois), chairman of the House appropriations subcommittee that oversees the NIH, last week called Clinton "contemptible" for publicly lauding the NIH and the importance of biomedical research without providing the dollars to pursue it. "The president has been great with the rhetoric, but he's been nowhere with the resources," Porter said.

Arlen Specter (Republican, Pennsylvania), Porter's Senate counterpart, took the director of NIH, Harold Varmus, to task at a hearing on 11 June. Varmus said that the administration had "assured" him that Clinton's February request for a 2.6 per cent (\$337 million) NIH increase still stood.

Specter called it an "illusion" that such an increase could be provided under the terms of the budget agreement, which would decrease by \$100 million the fund from which NIH and other health money is drawn. He called on Varmus and the other NIH direc-

tors to get to grips with these "hard facts".

But Lawrence Haas, an OMB spokesman, said last Saturday (14 June) that there is enough money in the budget agreement "not only in 1998 but over the next five years to fully protect the president's priorities, of which NIH has always been one".

Last Friday, Porter's subcommittee learned that the money available for its 1998 spending bill has increased by \$4.4 billion over 1997. Porter is an ardent advocate of biomedical research and could normally be expected to use part of this increase to boost the NIH budget substantially. In the 1997 spending bill, he increased NIH's funding by 6.9 per cent.

But the unusual circumstances of this year's budget agreement mean that Porter may find himself less able to direct new money to the biomedical research agency. The agreement protects, for instance, education — an area that is also part of the Porter subcommittee's purview. **Meredith Wadman**

Statistician is named to UK science post

[LONDON] The British government is announcing this week that Anne Campbell, the Member of Parliament for Cambridge, has been appointed to a post in the Department of Trade and Industry, where her responsibilities are to include various science-related issues.

John Battle, the minister for industry, energy and science, is said to have told department officials last week that Campbell was to be appointed as his parliamentary private secretary, a relatively junior post in the ministerial pecking order, but one that still holds the potential for wielding considerable influence.

Campbell's appointment is likely to be widely welcomed in the scientific community. She is a former chair of the Parliamentary and Scientific Committee, which brings together politicians with industrial and academic scientists. As an active member of the House of Commons Select Committee on Science and Technology, she has been a vocal critic of weaknesses in the previous government's science policy.

Shortly before the general election, Campbell said that she would like to see the position of the chief scientific adviser to the government strengthened to ensure that "the government's overall strategy for science is better followed" (see *Nature* 386, 315; 1997). She has also attacked the way in



Campbell: keen to see bolder strategy.

which the latest 'prior options' review of government-funded research laboratories was carried out.

Her appointment is expected to help assuage widespread concern in the scientific community that, although Battle is

formally identified as the 'minister for science' — a capacity in which he has already attended several meetings of European research ministers — his portfolio is too broad to allow him to give research a substantial amount of his time.

Campbell, who has been an MP since 1992, is a mathematician by training and a former college lecturer in statistics. In recent years, she has spoken out strongly in favour of increasing links between universities and industry, as well as calling for more government support for small, high-technology companies. She was also instrumental in helping to launch the group Scientists for Labour.

Campbell had been tipped by some as a possible science minister. But some feel that her prospects were tarnished when she declined to vote for Tony Blair in the Labour leadership election three years ago over the way that the vote was handled. **David Dickson**

Global fusion plans face three-year delay

[MUNICH] A decision about whether to build a multibillion dollar international facility to test the feasibility of nuclear fusion as an energy source is almost certain to be delayed by three years because of lack of political enthusiasm.

The four partners in the International Thermonuclear Experimental Reactor (ITER) project — the European Union (EU), Japan, United States and Russia — set up an 'exploratory' committee last year to look at non-scientific issues relating to the siting of the facility. The committee has concluded that, given the financial difficulties faced by all the partners, it is too soon to move to the construction phase when the current design phase ends in July next year. A formal decision by the partners on the proposed delay is expected early next year.

The conclusions of the committee coincide with new thinking about support for the project within the EU. The union's member states finance a joint fusion programme to which the European Commis-

sion has contributed nearly ECU900 million (US\$1 billion) during its five-year fourth Framework research programme (FP4).

The level of support for fusion research in the fifth Framework programme (FP5), due to succeed FP4 in 1999, is currently under discussion. If the construction of ITER were



Keilhacker: time to improve design?

to start next year, as originally planned, then the support would need to be increased in FP5 and future Framework programmes by between 10 and 40 per cent, depending on whether the facility is sited in Europe. Some EU member states, such as the United Kingdom, Sweden and Austria, strongly oppose any increase in the fusion budget, and France and Germany are unenthusiastic about the idea. The European Parliament, which has the final say in approving the FP5 budget, is deeply divided about support for fusion research on both

financial and environmental grounds.

Final financial decisions on FP5 must be made by early next year. To help decision-making, the Netherlands, which holds the EU's rotating presidency until the end of June, has asked the commission for its response to five options it has set out for ITER's future, including its comments on their implications for FP5 funding.

Two of the options are relatively extreme, and few take them seriously. One would be to stop fusion research completely. The second option would be to abandon the four-way ITER partnership and pursue a European-only version of ITER.

A further option would be to abandon plans to build ITER, but continue scientific research into fusion. But the Dutch paper warns that this option could cause the research to lose its focus on developing a source of future energy, because it would no longer be linked to the building of a reactor.

Another option, to start construction next year as originally planned, would mean increasing the fusion budget by 12 per cent in FP5 compared to FP4.

The option favoured by the Netherlands — and apparently by the commission — would be to postpone a decision on construction until after 2000, and to maintain the fusion budget at its FP4 level, extending funding for JET (the Joint European Torus) as a scientific bridge.

The predicted delay does not necessarily mean that ITER would come into operation later than planned, says Martin Keilhacker, director of JET. Instead, things would be done in a different order. A two- to three-year interim phase would allow a more detailed technological review of ITER to be completed, for example.

Keilhacker says that this would give time to carry out more scientific experiments to help reduce the uncertainty about the design, particularly the key issue of ignition of the fusion plasma.

ITER partners also suggest that this time could be used to make progress on choosing potential sites for the facility. Keilhacker says it would make sense for pre-licensing procedures to begin for several potential sites at the same time, because licensing is a factor in determining the rate of progress in construction of a reactor, which can take many years.

But companies involved in reactor building have expressed disappointment about the proposed delay. Matthias Kohler, executive director of Siemens's fusion programme, says that "as a taxpayer" he can understand the need for delay. But as a representative of industry he is "disappointed, because ITER represents the first chance for industry to begin to be a real player in the development of fusion technology".

Alison Abbott

Collapse of study complicates US participation

[WASHINGTON] An independent study that would have played a key role in determining the shape of US participation in the ITER project has collapsed, the first major victim of a legal impasse at the National Academy of Sciences over the status of its expert panels.

The United States will spend \$55 million on ITER this year. The Department of Energy (DoE) had commissioned the study to help to decide what to do after the ITER engineering design assessment ends in July 1998 (see above).

Perhaps more importantly, from the department's point of view, the study would probably have endorsed at least some of ITER's achievements, making it easier to win congressional support for continued US participation.

Earlier this year, the DoE asked the National Research Council — the operating arm of the National Academies of Science and Engineering — to set up a panel to assess the ITER design by October, in time to help prepare the

department's 1999 budget proposal for fusion energy sciences. But as a result of a recent court case the council and the DoE fear that they could face legal action under the Federal Advisory Committee Act if they appoint a panel in the usual way (see *Nature* 387, 220; 1997).

Lawyers have been trying to design a study that would avoid the act's rules by putting a single investigator in charge. But they abandoned their efforts after being told by congressional staff that the outcome of such a process was unlikely to have much credibility in the Congress. Martha Krebs, assistant energy secretary, confirms that, as a result, the study will not now proceed.

The DoE will now have to rely on its own Fusion Energy Science Advisory Committee (FESAC) to assess ITER and plot a way forward. The latest ITER design was endorsed in April by a FESAC panel chaired by Robert Conn, dean of engineering at the University of California at San Diego (see *Nature* 386, 745; 1997).

Another panel, led by Herman Grunder, director of the Thomas Jefferson National Accelerator Facility in Newport News, Virginia, is now being asked to produce some recommendations for US involvement in an extended ITER design assessment.

Krebs is pragmatic about the setback. The academy review would have been helpful, she says, but advice from FESAC will have to do instead. "I believe we can make a strong case that the activities we'd do [under an extended design assessment] would be science-driven and would sustain our domestic capability in fusion."

But, despite a more positive attitude towards fusion research in the Congress than in the past, selling future US involvement in ITER, already a tough task, is likely to become still more difficult. Although FESAC has some scientists from outside fusion research, the committee is seen primarily as the voice of the fusion community. **Colin Macilwain**

Tritium leak protests prompt Canada to suspend discharges

[MONTREAL] The Canadian government has ordered a halt to the disposal of liquid waste from Atomic Energy of Canada's main nuclear research site, following protests from environmental groups provoked by details of leaks of radioactive waste water.

According to Canada's Atomic Energy Control Board (AECB), tritium levels in the leaks at Chalk River, Ontario, are well within regulatory standards and are not a danger to public health — despite the fact that some of the waste has entered the Ottawa River, whose waters are used for drinking.

The situation echoes that at Brookhaven National Laboratory, New York, whose director resigned and whose long-time contractor was sacked following tritium leaks (see *Nature* 387, 114; 1997). In both cases, government action seems to have been prompted by public concern rather than by evidence of a clear danger.

The leakage, which has taken place over a 20-year period, was from uranium fuel storage bays for the now inoperative NRX research reactor, and put thousands of litres daily into the river. Atomic Energy of Canada Ltd (AECL) made studies over the years and reported them in annual reports to the AECB but did not otherwise go public with them.

The warden (the senior elected official) of Renfrew County, where the Chalk River plant is situated, has complained that his council was never told of the leak. Others who feel the council should have been informed include Lynn Jones of nearby Pembroke, who uncovered the extent of the leak via a request filed under Canada's 'access to information' law.

AECL points out that it reported the leak in publicly available annual reports, as it was required to do, but did not want to alarm the public unnecessarily by further disclosures.

Two more leaks were revealed later, in pipes carrying water from routine drainage of reactor liquids into a gravel-filled pit on AECL property. The AECB says that the total amount of tritium in the leaks corresponds to only 0.01 per cent of the legally permissible limits on releases from the site.

Public concern about the leaks is linked to an offer by the nearby town of Deep River, which has since been dropped, to accept radioactively contaminated soil for storage at Chalk River. The soil is from a uranium processing plant near Port Hope, Ontario. Opponents of the plan questioned whether Chalk River is capable of storing the soil if it cannot adequately control its own waste.

Robert Potvin, an AECB spokesman, admits that AECL "should have been a little more open to local communities". **David Spurgeon**

Space physicists plead for extended solar missions

[WASHINGTON] Intense budget pressures are forcing a cash-strapped US National Aeronautics and Space Administration (NASA) to make the difficult decision about which of 10 current solar-terrestrial spacecraft missions should continue to operate beyond their planned lifetimes.

Among those facing possible shutdown in the next 18 months are POLAR, WIND, GEOTAIL and the European-built Solar and Heliospheric Observatory (SOHO). These form the core of the \$2.4-billion International Solar-Terrestrial Physics Program (ISTP) which is studying Sun-Earth interactions from different vantage points in the inner Solar System.

Earlier this year, NASA invited those involved in ISTP and other projects, such as Ulysses and SAMPEX, to propose plans for extended missions. Following three days of presentations last week, a panel chaired by Lennard Fisk will rank the proposals by scientific merit and report back to NASA this summer. Fisk is the former NASA science chief, and is now a space physicist at the University of Michigan.

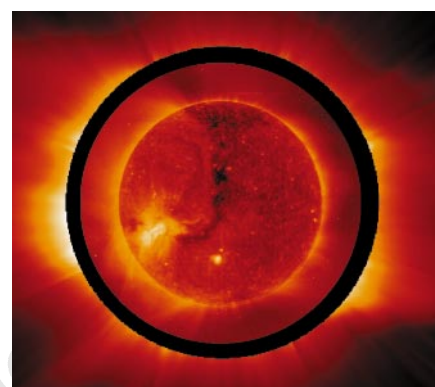
The problem facing the panel is similar to one presented by astrophysics missions last year (see *Nature* 383, 747; 1996), namely what to do with spacecraft that overstay their budgetary welcomes. Although the ISTP network was only completed last year, NASA funds for operating the spacecraft and analysing their data will run out by the end of the 1998 fiscal year.

"In our case, the welcome was very short," says Mario Acuña, ISTP project scientist at NASA's Goddard Space Flight Center in Maryland. WIND and POLAR were granted only three years of operational life. SOHO, which was launched in December 1995, got only two years.

That was not the plan sold to the scientific community, says Glenn Mason of the University of Maryland, who heads NASA's advisory committee for 'Sun-Earth connections'. He points out that the network was originally intended to remain in full operation until the next solar maximum, which occurs around 2001.

According to Mason, Daniel Goldin, the NASA administrator, raided 'out-year' funding for mission operations and data analysis in 1994 to improve his long-term budget picture, leaving the ISTP with no money after 1998. The funds were never restored.

Acuña and his team last week presented a plan to extend missions through the solar maximum at reduced cost by updating computers, using off-the-shelf software and reducing monitoring activities. Operations



Burning issue: SOHO, source of this ultraviolet image of the Sun, could be shut down.

teams for some of the ISTP spacecraft would be merged, and investigator teams would be scaled down and asked to compete with the rest of the scientific community for grants to analyse the spacecraft data.

Guenter Riegler of NASA's Office of Space Science says that the budget for data analysis tends to be "more resistant to shrinking" than mission operations, as it pays directly for scientists' time. Acuña says the research programme for the extended WIND, POLAR and GEOTAIL missions would cost only about \$25 million up to 2001, roughly half current levels.

But NASA's dilemma is that it cannot afford even that. Wesley Huntress, head of science at the space agency, challenged the Goddard centre, which also runs the Hubble Space Telescope and other science missions, to cut its operations costs across the board to help pay for extended missions.

But Riegler admits that even though Goddard has identified "a fair amount of money" that might be saved by squeezing parts of its operations, it is "not anywhere near" the amount needed to fund the ISTP missions to 2001.

Mason hopes that Congress will come to the rescue, and that a recent decision by the House of Representatives to add money to NASA's authorization bill specifically for data analysis will survive the budget appropriations process. Acuña says that whatever advice comes out of the latest review, "they're not going to print money".

Some scientists say that NASA's top managers, as well as politicians, also need to understand that it can take several years to get the most out of a spacecraft mission, and that simply posting data quickly on the Internet does not guarantee a high scientific yield. Careful analysis takes time and consistent funding for research, says Mason. "It's not a matter of just picking up nuggets on the beach."

Tony Reichhardt

Republicans seek to widen cloning ban

[WASHINGTON] Almost as soon as President Bill Clinton's proposed law to outlaw human cloning reached the US Congress last week, it became evident that it does not go far enough for many Republicans.

Conservative Republicans in the House of Representatives and the Senate promised to push for a tougher law to ban permanently not only human cloning for the purposes of reproduction, but also the use of cloning in the private sector to create research embryos not intended for implantation.

Clinton's Cloning Prohibition Act has been drafted in response to the recommendations of the president's National Bioethics Advisory Commission (NBAC). It would ban for five years "any attempt to create a human being using somatic cell nuclear transfer cloning", the method used by Scottish scientists who announced in February that they had cloned Dolly the sheep (see *Nature* 385, 810–813; 1997.)

The act would impose a fine of at least \$250,000 on anyone caught attempting to do so, and additional profits would be liable to seizure. The NBAC would be required to recommend whether the law should be renewed as its expiry date approached.

The bill has not been introduced in Congress, a procedure that requires a member of the House or Senate to sponsor it. None has stepped forward to do so, although it was not clear last week whether this is due to reticence or to simple procedural delays.

But, at hearings in the House and Senate last week, two politicians who have already introduced anti-cloning bills of their own complained that the Clinton measure — and the NBAC recommendations on which it is founded — are morally inadequate.

Senator Christopher Bond (Republican, Missouri) complained that the NBAC had avoided a politically charged issue by remaining silent about embryos created by cloning but not implanted. He said: "By allowing cloning research on human embryos in the private sector, the commission said: 'Go ahead as far as you can; when it gets dangerous then we'll try and stop you'."

Bond said this approach — also implicit in the Clinton bill — "risks sliding very far down the slope" to human cloning and that tighter controls were necessary.

Both Bond and Congressman Vern Ehlers (Republican, Michigan) insisted that any legal ban should be permanent. Ehlers, a physicist by training, has introduced two bills — one to outlaw cloning for the purposes of reproduction, and the other to outlaw research toward this goal.

Ehlers said that the House would not stay silent on whether to allow private production of cloned embryos for research.

He suggested that any debate would require the House to discuss a broader question — why human embryo research should be allowed in the private sector at all. Current US law bans federally funded

human embryo research in which embryos are harmed or destroyed, but such research is allowed in the private sector.

At the House hearing, even liberal politicians questioned the consistency of the NBAC recommendation. Sheila Jackson Lee (Democrat, Texas) declared that the question of unrestricted activity in the private sector "looms".

"How do you [rationalize] that bifurcated" recommendation, she asked Harold Shapiro, the NBAC chairman and president of Princeton University. Shapiro responded that different standards apply when contentious moral questions relate to taxpayer-financed activities as opposed to private ones.

George Annas, professor of law and public health at Boston University, suggested that, for anti-abortionists, the Clinton bill is "the worst of both worlds", as it allows the creation of research embryos in the private sector, and requires their destruction.

Annas speculated that Congress was more likely to prohibit all research on human embryos than explicitly to sanction research on human embryos with no chance of implantation. He also said that Congress was unlikely to be content with the civil penalties in the Clinton bill.

"The president's [bill says] it's wrong to do this because it puts the [resulting] children at a very grave risk of genetic harm," Annas said. "I think Congress will determine this is a felony."

Meredith Wadman

Palaeontologists protest at Web sale of hominid remains

[LONDON] An undergraduate palaeontology student from Texas, who caused controversy last month by trying to sell a *Tyrannosaurus rex* skeleton through the World Wide Web, has caused a further stir by selling a hominid skull and jawbones to an undisclosed university in Europe.

Jim Wyatt, a student at the University of Texas, Dallas, owns Fossilnet, an electronic 'supermarket' of fossils. Earlier this month, he put pictures on his website of what are believed to be Cro-Magnon and Neanderthal human remains from Italy, including a skull (right) valued at US\$28,000.

The European university, which has a large department of palaeoanthropology, is believed to have struck a deal at the start of the month. The skull and seven other items were jointly priced at \$94,000. But Wyatt says the university paid "substantially below the asking price".

The items are part of a private collection that once belonged to an Italian fossil hunter, Frederic Zambelli Hosmer. They are believed to have been excavated in the 1920s and '30s, and sold by Zambelli to



finance his explorations.

Last month, Wyatt angered members of the Society for Vertebrate Paleontology by using as the showpiece item of his website a virtually complete — and still unsold — privately owned skeleton of *T. rex* with a price of \$12 million.

His decision to sell human remains "is just going too far", says Eric Scott, palaeontology field supervisor at the San Bernardino County Museum in California. "Many people have objected to the sale of

dinosaur skeletons. But selling human material is just frightening. And there's so little of it around."

Wyatt claims to earn less than \$50,000 a year, mostly by brokering small-scale deals between private collectors and museums and universities. "What I'm doing is totally legal, and no different to any other business," he says.

Buying and selling fossils brought legally from another country is permitted in the United States, as is fossil collection on private land. A proposed bill — the Fossil Preservation Act of 1996 — would even permit commercial fossil collection on public land.

The Society for Vertebrate Paleontology is opposed to this, as is the public. A clear majority of respondents to a survey conducted two years ago said they believed fossils found on public or private land should belong to public institutions (see *Nature* 379, 388; 1996).

An associated organization, Save America's Fossils for Everyone (SAFE), has drafted its own fossil-preservation proposals as a draft Senate bill.

Ehsan Masood

Brussels reshapes science panels after BSE criticism

[PARIS] As part of an overhaul of its system for gathering scientific advice on food and health issues, the European Commission in Brussels last week announced details of how it plans to reorganize its advisory committees.

It hopes these changes will help to restore public confidence in an advisory system whose image has been badly tarnished by the commission's failure to prevent the crisis over bovine spongiform encephalopathy (BSE). Advisory committees, and the commission itself, have been criticized for allegedly paying too much attention to national and economic interests to the detriment of consumer safety (see *Nature* 385, 6; 1997).

The reorganization of the advisory committees is similar to that promised earlier this year by the commission's president, Jacques Santer (see *Nature* 385, 664; 1997). The six committees covering food and public health have been moved from the industry and agriculture directorates to the consumer directorate, under a scientific steering committee that will coordinate their work.

The steering committee will be chaired by

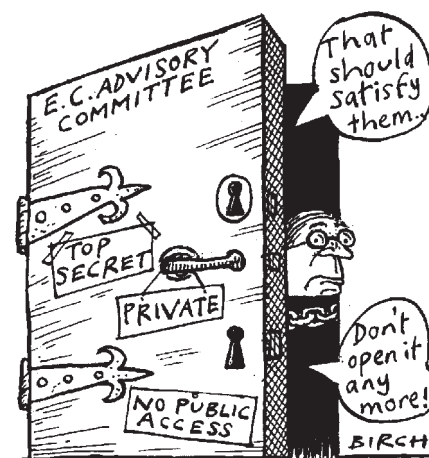
Fritz Kemper, a virologist from the University of Münster in Germany, and will include the chairpersons of the six advisory committees and eight outside experts.

Minutes of advisory committee meetings, including details of minority positions, will in future be made public, a measure broadly welcomed by consumer organizations. But these organizations are unhappy that responsibility for setting policy and proposing legislation on food safety will remain with the agriculture and industry directorates, which also look after the interests of industry and agriculture.

Jim Murray, director of the European federation of consumer organizations (BEUC), in a letter to Santer, argues that "without a clear separation of [these] functions, conflicts of interest will continue to arise".

Rachel Kenningham, an official with the UK Consumer Association, says that a better system would have been to give executive powers to the consumer directorate that would have allowed it to propose legislation and codes of practice.

But Kenningham, like many other observers, feels that the changes, which will



take effect next month, represent significant progress. They argue that allocating responsibility for scientific advice to a strengthened consumer directorate, combined with greater openness and independence of advisory committees, will expose decisions on scientific issues to more public scrutiny. "At least the decision-making will now be in daylight and not behind closed doors," says one advisory committee member.

One major change concerns the nomination of experts to the committees. Until now the selection procedure has been opaque, leading to questions about the competence and independence of nominees, and the criteria on which they were appointed.

The commission now intends to recruit experts through an open competition, or concours. Candidates will be assessed by a selection panel of commission officials chaired by an eminent scientist, with observers from the European Parliament and the Council of Ministers present.

The commission will then nominate experts from the panel's shortlist. All candidates will be required to declare any potential conflicts of interest.

But there could be a drawback. One commission official argues that eminent scientists often feel that it is "beneath" them to have to apply for appointments through a procedure akin to that used for recruiting junior civil servants. Several have already complained, according to the official, who added that as a result the commission risks recruiting "mediocre" experts.

The European Parliament has warned that if the commission has not implemented sufficient reforms by the end of the year, it will consider using its powers to vote a motion of censure. All the commissioners would then be dismissed.

The BSE crisis might have been avoided if there had been greater openness in its handling in Brussels, says one member of the commission's scientific veterinary committee. He claims that the 30 or so meetings of the committee on BSE produced a mass of information and debate, little of which reached the public domain.

Declan Butler

Britain extends controls to combat risk to sheep

[PARIS] The British government last week announced that it is to extend strict bovine spongiform encephalopathy (BSE) controls to sheep, to guard against the risk that the disease has jumped to sheep and is coexisting in flocks alongside scrapie.

The new measures, which were announced by the agriculture minister, Jack Cunningham, include removal in the abattoir of the spinal cord from sheep and goats over one year of age, and the spleen from all animals – heads are already banned from human consumption and animal feed.

The risk that sheep have acquired BSE from contaminated animal feed was raised by the government's Spongiform Encephalopathy Advisory Committee. Oral transmission of BSE to sheep has been demonstrated experimentally, despite no evidence for its

existence in the field.

Cunningham said the government was taking "sound, precautionary measures" to avoid any possible risk to consumers, "no matter how remote". Similar controls have been introduced in France, Ireland and the Netherlands.

But European Commission proposals for controls on sheep throughout the European Union, as recommended by its scientific veterinary committee, have been repeatedly rejected by the council of farm ministers, with several countries arguing that they do not have BSE (see *Nature* 384, 8; 1996).

In a separate move, the British government is threatening to ban the import of beef from countries with BSE unless they bring their controls up to UK standards.

The European Commission says such a unilateral move would be

illegal. But Britain is finding sympathy in some quarters. "Cunningham's move is extremely sensible," says one European Parliament official.

Britain is also likely to find ammunition for its demands in a paper submitted to the *Veterinary Record* suggesting that BSE has been under-reported in countries on the continent, and that the approximately 50 declared cases are fewer than the several thousand expected, given the level of exports of contaminated animal feed and cattle (see *Nature* 382, 4; 1996).

The findings, from an international research group led by Bram Schreuder from the Dutch DLO-Institute for Animal Science and Health in Lelystad, are likely to weaken the position of Germany and other countries who oppose stricter BSE controls on the grounds that they have no BSE.

D. B.

Fraud claims shake German complacency

[MUNICH] The research results of two prominent molecular biologists were systematically fabricated over a number of years, according to a panel of scientific and legal experts which met in Bonn last week.

Accusations of fraud, which have sent shock waves through Germany's scientific community, have been made against Marion Brach, from the University of Lübeck in the state of Schleswig-Holstein, and her former colleague Friedhelm Herrmann, from the University of Ulm in Baden-Württemberg (see *Nature* 387, 442; 1997).

They have already prompted moves in the research community to reassess the procedures for handling such cases, and in particular to ensure that such accusations are investigated fairly and rapidly.

The alleged fraud is said to have taken place when Herrmann and Brach worked together at the Max Delbrück Centre for Molecular Medicine (MDC) in Berlin in the early 1990s. But the allegations were made only earlier this year by young co-researchers who had become suspicious.

The MDC and the two universities immediately set up committees to investigate the case. The Bonn committee, set up by the heads of the local committees, considered the evidence from these investigations. It also took further evidence from Brach, who has admitted fabricating results in four papers published between 1994 and 1996, and from other witnesses. Herrmann denies all charges and declined to attend the Bonn meeting.

The Bonn committee confirmed that each

of the four papers included falsified data, and also heard evidence on 20 other published papers, which will be investigated further. The committee is now preparing a detailed report to which Herrmann and Brach will be invited to respond before it is published.

Despite the committee's conclusions, the universities cannot dismiss the two researchers, as they are employees of the local state (*Land*). Horst Laqua, dean of medicine at the University of Lübeck, has therefore asked the ministry of education and research in Schleswig-Holstein to dismiss Brach on the basis of her confession.

But even this may not be straightforward. Laqua points out that because none of the fraudulent activities took place at Lübeck, to which Brach has moved only recently, the legal position is unclear.

The Baden-Württemberg ministry of research will find it even harder to take conclusive action against Herrmann. He continues to deny all charges against him, and has rejected requests from the faculty to stop working voluntarily until the case is resolved.

Action may be made easier by the outcome of an investigation requested by the public prosecutor in the state of Baden-Württemberg, who wants to establish whether grant money given to Herrmann and Brach was gained fraudulently or used for fraudulent purposes.

The Deutsche Krebshilfe Stiftung, a cancer research charity that has given grants of around DM1 million (US\$580,000) to Herrmann and Brach, has said that if the case

is proved it will seek the return of the money.

The Deutsche Forschungsgemeinschaft (DFG), Germany's major university grant-giving body, is also studying whether any of its grants were used in the fraudulent work.

The case has shaken complacency about scientific integrity in Germany, where few guidelines exist to handle such cases. The DFG is one of the very few organizations to have its own rules about fraud investigations. These include the ultimate — and as yet untested — sanction of demanding the return of grant money.

The Max Planck Society (MPS) set up a committee last year to establish its own procedures for handling fraud. Its chairman, Albin Eser, director of the Max Planck Institute for International Criminal Law, says that it is difficult to devise procedures that are compatible with German law.

But Eser stresses the need for a system that will protect the whistle-blower — the scientists who raised questions about Herrmann and Brach were worried that their actions could adversely affect their own careers — while allowing those accused to defend themselves against false accusations.

It is also important that someone found guilty of fraud is not destroyed for life, says Eser. As an example of good practice, he cites the case of a German scientist working in Switzerland who fabricated results in papers published with scientists at the Max Planck Institute for Psychiatry in Munich and elsewhere. The case was reported last week in the news magazine *Focus*.

One of these scientists had concluded that work cited in a paper by the Zurich-based scientist must have been invented. The scientist immediately admitted routinely fabricating results that had been published with collaborators throughout Europe.

He was instantly dismissed by Zurich University, and papers including his fraudulent work were withdrawn. This case received no publicity at the time, and he is now working as a medical doctor in Switzerland.

Eser has long been campaigning for all research institutions in Germany to prepare procedures for handling fraud. In the past, he says, the scientific community has tried to play down the incidence of fraud, and has tended to believe in its own "self-healing powers". The latest case has forced a change in attitude because of its audacity, he says.

Hubert Markl, the MPS president, says that all research institutions are discussing how they should react. But he says that there is unlikely to be support for setting up a central office for scientific integrity, or a central set of guidelines for use throughout the country, as each research organization and each university likes to protect its independence.

Alison Abbott

University 'failed to acknowledge exoneration'

[MUNICH] A former professor of developmental biology at the University of Geneva, where he was accused of fraud in the early 1980s, has complained that the university has for years refused to acknowledge his exoneration (see *Nature* 307, 673; 1984).

The University of Geneva set up an international commission of experts to investigate allegations made by his younger co-workers that Karl Illmensee fabricated (unpublished) data in nuclear-transfer experiments; such experiments are important in the development of mammalian cloning technology. The commission found no evidence in his research protocols to support or refute the accusation, and

suggested that the experiments be repeated in an international collaboration.

This suggestion was followed through by Illmensee and results were published in the peer-reviewed journals *Naturwissenschaft* (1989) and *Development* (1990). Two members of the commission, Richard Gardner, a professor at Oxford University, and Anne McLaren, a principal scientist at the Wellcome Research Institute in Cambridge, wrote to the University of Geneva in 1991 explaining that the papers reproduced the "essential findings" in Illmensee's earlier experiments, despite the use in later experiments of a different cell line for practical reasons. "We consider it

appropriate that the University of Geneva should inform the scientific community that these controversial findings have now been confirmed under the conditions specified by the commission," they wrote. They received no answer.

Illmensee says he left Geneva because of the bad feeling generated by the affair and now works at the University of Salzburg.

The current rector of the University of Geneva, Bernard Fulpius, says that as he has been in office for only two years, he is not familiar with the details of the affair. But he says that he plans to reply to Illmensee's most recent letter, which was sent in May, later this month.

A.A.

Nuclear triggers get subcritical scrutiny

Colin Macilwain

Experiments in Nevada will not only increase our understanding of nuclear explosions, but should also help secure the future of the test site.

Five years after the last of 880 nuclear weapons tests shook the Nevada Test Site, scientists are preparing experiments there involving small underground explosions that they hope will yield useful information about the behaviour of the plutonium triggers in existing nuclear weapons.

Like the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory in California, these so-called subcritical experiments will feed information into the new nuclear weapons codes being developed under the US Science-based Stockpile Stewardship (SBSS) programme (see *Nature* 387, 541; 1997).

And also like NIF, the experiments have a political purpose: they will give Nevada a share in the SBSS programme and reassure local politicians that the end of testing will not mean the end of the test site, whose operation brings more than \$300 million of federal money into Nevada each year.

The subcritical experiments may lack NIF's lustre, but they have at least shared in the controversy that surrounds it. The first such experiment, called Rebound and conceived at Los Alamos, is now ready. But, like NIF construction, it is being held back until at least 27 June by court action against the SBSS programme.

Advocates of the subcritical experiments — which produce no measurable fission yield — say that they are vital to stockpile stewardship and also in keeping with both the spirit and the letter of the Comprehensive Test Ban Treaty. They worry that the public confuses them with two other classes of experiment, now abandoned by the United States — hydronuclear experiments and low-yield nuclear tests, which could respectively produce fission yields equivalent to 2 kg and 250 kg of high explosive.

Activists' critical questions

Critics — including local activists who were arrested trying to block a coachload of journalists taken to the site by the energy department last month — ask why, if the experiments are so harmless, they are being conducted in secure caverns 960 feet below the desert.

But according to Robin Staffin, deputy assistant secretary for weapons research at the Department of Energy (DoE), foreign governments are satisfied that they fall with-

in the remit of the test ban treaty, and have not asked to send observers.

The Rebound experiment will ignite three blocks of 15, 25 and 40 kg of high explosives next to plastic rigs containing several matchbox-sized blocks of plutonium. Sensors will optically measure the speeds of shock and sound waves through the samples, providing equation-of-state data on plutonium, for use in the new weapons codes.

The total amount of plutonium in the chamber will be 1.5 kg — not far short of the secret amount needed to produce a fission reaction under ideal conditions. The plutonium is, however, divided into more than a dozen samples, and a neutron detector on the roof is expected to confirm that there is no detectable fission yield from Rebound.

The next subcritical experiment planned at Nevada is a Livermore one called Holog, which will use holography to photograph the tiny fragments that fly off the surface of a plutonium coin when it is punched with 50 grams of detonating high explosive. Asked what the experiment is for, Mike Dunning, one of its designers, gives an old-fashioned answer not quite in keeping with the new spirit of openness at the DoE: "I can't go into

the programmatic reasons why we're doing this," he says. "The fact that we're doing it ought to tell you that it matters."

But the surface behaviour of plutonium is a critical factor in the operation of any nuclear weapon. The Livermore team, explains Dunning's colleague Dick Iear, wants to know how it will change as plutonium ages: in 40 years, one plutonium atom in a thousand will decay, changing its already complex properties. Although they do not say so, they may also want to know how surface behaviour will be affected if machined weapons pits are remanufactured as castings.

The energy department cites a 1995 study by the Jason group of scientists as independent evidence that these experiments are worthwhile. Sidney Drell, deputy director of the Stanford Linear Accelerator Center, who led the study, says that "the knowledge gained from them is going to be critical" to the maintenance of the stockpile.

Economic role for Nevada

They are also important to the economy of Nevada. They will be used as rehearsals to keep the test site prepared for any resumption of weapons testing, and will become the main function of a piece of desert that costs more than \$300 million a year to run.

"I'm still not convinced that they are needed for the maintenance of the stockpile," says Suzanne Jones, a physicist at Frank von Hippel's Center for Energy and Environmental Studies at Princeton University, New Jersey. "I think they want to appease the laboratories, which need stuff to do, and the Nevada senators. They think if they don't do the experiments, the Senate won't ratify the Comprehensive Test Ban Treaty." □



Going underground: the Nevada test site will host experiments to gather data on nuclear triggers.

Canadian ministers send mixed message on research funds

[MONTREAL] Canada's scientists have been receiving different signals from the new science minister and the minister of industry about whether the federal Liberal government, re-elected last week, will restore the budgets of the research councils.

The councils are Canada's main source of research funds and have suffered a 15 per cent cut over the past three years. Science minister Ronald Duhamel, a former university lecturer, is reported to have said shortly after the election that any commitment to restoring the cuts would be premature.

But John Manley, the industry minister to whom Duhamel reports, and who is one of seven senior cabinet ministers who kept their portfolios, promised to fight for the research councils as priorities in his second term.

US panel targets cuts in nuclear weapons

[WASHINGTON] The United States and Russia should aim to reduce their nuclear weapons stockpiles to 1,000 warheads each, with a view eventually to "prohibiting" the weapons altogether, according to a panel of the US National Academy of Sciences. The new

limits would be almost an order of magnitude stricter than those proposed under the still-to-be-negotiated Start III treaty, which will allow up to 2,500 warheads to be deployed at any one time — with thousands more held in reserve and in various states of maintenance.

A deal covering all warheads would require far more intrusive verification procedures, says the panel, which was chaired by William Burns, a retired army general. A 1991 study by the academy helped to lay the basis for the Start II treaty and the Start III negotiations, but did not propose elimination of the weapons.

Israeli university denies 'illegal' computing

[LONDON] The president of Israel's Ben Gurion University has issued a statement denying media reports that the university's Cray supercomputer is being used for "illegal" nuclear weapons research. The reports had said that the US government had banned US companies from doing business with the university, believing it to be involved in a joint project with the Dimona nuclear facility in southern Israel.

The reports claimed that US authorities had placed the university's name in the *Federal Register* along with other organizations believed to be involved in the

illegal use of American technology. But US government officials deny the existence of any ban. They say that the university was listed precisely because it is an "approved and legitimate user" of the supercomputer, and the names of such users need to be made public under US law.

Russian suicide blamed on economic problems

[MOSCOW] A leading Russian geophysicist who acted as a consultant to the State Duma, the Russian parliament, has apparently committed suicide because of growing economic difficulties at his institute.

Evgeny Isaev, director general of the Zarubezhgeologia scientific research institute, who was found dead in his office, left a lengthy letter addressed to the tax office and the department of economic crimes detection. In the note, Isaev is said to have complained that employees had not received salaries since March. Isaev was the author of more than 130 scientific works, and held five patents. He joined the institute in 1985.

Brookhaven names temporary director

[WASHINGTON] Associated Universities Inc. (AUI), the embattled contractor at Brookhaven National Laboratory, New York,

has nominated David Moncton, a former Brookhaven scientist, to be director of the laboratory until AUI's contract ends in November. Moncton, who runs the Advanced Photon Source at the Argonne National Laboratory in Illinois, is still negotiating with his employer about the temporary move, which also must be approved by the Department of Energy. Lyle Schwartz, the interim director at Brookhaven, has been criticized by some scientists at the laboratory who say that he cannot lead them effectively while also serving as president of AUI.

Call for Europe's research plan to stress health

[MUNICH] **The European Commission's proposed fifth Framework programme (FP5) should be revised to include a thematic programme dedicated to health research, according to a new report from the European Science Foundation's standing committee for the medical sciences.**

The committee, which represents more than 30 European biomedical research funding agencies, says that health issues could get lost within the proposed programme, called "Unlocking the Resources of the Living World and Ecosystem", because of its breadth and heterogeneity. To capitalize fully on Europe's biomedical research

potential, it argues, health requires its own programme, with clearly defined key actions, "The Ageing Population", for example.

Fermi nuclear awards announced by Clinton

[WASHINGTON] President Bill Clinton announced last week that Richard Garwin, Mortimer Elkind and H. Rodney Withers have won the Enrico Fermi award, a US government award for achievements in nuclear energy.

Garwin, a physicist with IBM, receives the \$100,000 prize for his scientific work and his advice to the government on nuclear weapons issues. Elkind, a cell biologist at Colorado State University, and Withers, an oncologist at the University of California at Los Angeles, share a further \$100,000 for their work on the response of cells to radiation, which laid the basis for radiation therapy of cancer. Federico Peña, the energy secretary, will present the awards on 24 July.

Glaxo-Wellcome sets up genetics directorate

[LONDON] **The pharmaceutical company Glaxo-Wellcome has set up a new genetics directorate to boost the role of genetics in drug discovery. It will be headed by Allen Roses, director of the Center for Human**

Genetics at Duke University, North Carolina, and widely known for his work on the genetics of Alzheimer's disease.

The directorate will have a budget of more than £30 million (US\$49 million) for 1997. Staff numbers are expected to double over the next 18 months. The company had earlier announced a US\$9-million acquisition of Spectra Biomedical in Menlo Park, California, a pioneer in the field of 'association genetics', a means of establishing associations between genes and specific diseases.

Physicists requisition café society in Paris

[PARIS] The traditional Paris hangout of French philosophers and intellectuals, the Boulevard Saint-Germain, will be invaded by physicists for three summer evenings next month when the Société Française de Physique requisitions the Vagenende 1900, a stylish Art Nouveau café on the boulevard, as a venue for public debate on physics.

The 'Bar des Sciences', as the café will be christened for the three days, is being organized as a fringe event for the 1,000 delegates at the society's annual conference, which will be held from 7 to 10 July at the Sorbonne. The aim, say the organizers, is to allow researchers to debate physics with "ordinary people in an ordinary place".

Is cloning an attack on human dignity?

Sir — Appeals to human dignity, and to the moral obligation to protect it, have been a feature of responses to the cloning of Dolly the sheep (*Nature* **385**, 810–813; 1997). Dr Hiroshi Nakajima, director general of the World Health Organization (WHO), said: “WHO considers the use of cloning for the replication of human individuals to be ethically unacceptable as it would violate some of the basic principles which govern medically assisted procreation. These include respect for the dignity of the human being.”

The European Parliament rushed through a resolution banning cloning, which stated as part of the rationale (clause 6) that the parliament “believes it is essential to establish ethical standards, based on respect for human dignity, in the areas of biology, biotechnology and medicine”. Neither of these august authorities provided a scrap of argument as to how the idea of human dignity is relevant to the ethics of cloning.

A first question to ask when the idea of human dignity is invoked is: whose dignity is attacked and how? Is it the duplication of a large part of the genome that is supposed to constitute the attack on human dignity? If so, we might legitimately ask whether and how the dignity of a natural twin is threatened by the existence of the other twin. The notion of human dignity is often also linked to Kantian ethics. A typical example, and one that attempts to provide

some basis for objections to cloning based on human dignity, is Axel Kahn’s invocation of this principle in his Commentary on cloning (*Nature* **386**, 119; 1997).

But the Kantian principle, which is generally interpreted as demanding that “an individual should never be thought of as a means but always also as an end”, crudely invoked, as it usually is, without any qualification or gloss, is seldom helpful in a medical or bioscience context. It would outlaw blood transfusions and abortions carried out to protect the life or health of the mother. It would also outlaw one form of cloning, embryo splitting, which could allow genetic and other screening by embryo biopsy. One embryo could be tested to ascertain the health and genetic status of the remaining clones, and then destroyed. To this it is objected, *pace* Kahn, that one twin would be destroyed for the sake of another.

It is bizarre and misleading to marshal the Kantian principle as an objection either to using cell mass division to create clones for screening purposes, or to creating clones by nuclear substitution to generate spare cell lines. It is surely ethically dubious to object to one embryo being killed for the sake of another, but not to object to it being killed for nothing. In *in vitro* fertilization (IVF), for example, it is, in the United Kingdom, regarded as good practice to store spare embryos for future use by the mother or for disposal at her direction, either to other women who require donor

eggs, or for research, or simply to be destroyed. It cannot be morally worse to use an embryo to provide information about its sibling than to use it for more abstract research or simply to destroy it. If it is permissible to use early embryos for research or to destroy them, their use in genetic and other health testing is surely also permissible. The same would surely go for their use in creating cell lines for therapeutic purposes.

A moral principle that has at least as much intuitive force as that recommended by Kant is that it is better to do some good than to do no good. It cannot, from the ethical point of view, be better or more moral to waste human material that could be used for therapeutic purposes than to use it to do good. If it is right to use embryos for research or therapy then it is surely also right to produce them for such purposes, as is usual in IVF. Kant’s prohibition does after all refer principally to use. Of course some will think that the embryo is a full member of the moral community with all the rights and protections possessed by Kant himself. Although this is a tenable position, it cannot consistently be held by any society that permits abortion, post-coital contraception or research with human embryos.

John Harris

*Centre for Social Ethics and Policy,
University of Manchester,
Manchester M13 9PL, UK
e-mail: Rebecca.Bennett@man.ac.uk*

Protect patients’ rights

Sir — Those of us who are concerned about the increasing incidence of biomedical discrimination and violations of medical privacy appreciate your attention to these issues (*Nature* **386**, 533; 1997). We need, however, to deal directly with the fact that research scientists do not have an exemplary track record.

For example, recent media coverage of the Tuskegee experiment (see *Nature* **387**, 116; 1997) serves as a reminder of why the United States had to develop the institution of Human Subjects Review Boards. Abuses of the rights of patients have contributed to the increased development of notions of appropriate informed consent.

In this regard, I believe your coverage has put too much emphasis on the needs of research scientists, while underplaying the integrity and dignity of research subjects. For example, as the increasing number of state legislative enactments protecting genetic information indicates, the donor or source of the material must have control

over the sample and any access to it. As far as protecting the rights of the donor is concerned, it is immaterial whether the third party seeking access to the sample is wearing a lab coat or not.

Philip L. Bereano

*(Chair, American Civil Liberties Committee on
Data Access, Storage, and Dissemination)
Department of Technical Communication,
University of Washington,
Box 352195, Seattle,
Washington 98195-2195, USA*

Hubble ‘worth the price’

Sir — David Leverington claims to be measuring cost-effectiveness of astronomical observations, and your provocative headline, “Star-gazing funds should come down to Earth,” shows that his conclusion can lead to important consequences (*Nature* **387**, 12; 1997). But, as his Correspondence shows, he is really merely measuring the number of times various papers have been cited.

He has no way of accounting for the uniqueness of an observation or considering the changes in our understanding that can result from breakthrough discoveries. His reliance on citation is like poll-driven government: it has some value but isn’t necessarily the best.

In particular, he says that, “although the Hubble Space Telescope has been extremely successful, Figure 1 shows that it has not yet justified its high costs...”. I submit that this conclusion is ridiculous and that, after seeing the wealth of unique and revealing observations that have been and are still pouring from Hubble, an observational scientist is driven to finding out what is wrong with Leverington’s arguments rather than accepting his conclusion. After all, if we assign a factor of 3 for uniqueness, then the space observations would meet even his absurd criteria.

Jay M. Pasachoff

*Hopkins Observatory,
Williams College,
Williamstown, Massachusetts 01267, USA
e-mail: jay.m.pasachoff@williams.edu*

A laser that sings a different tune

Sean Washburn

A new type of cascade laser has been built that can be tuned in wavelength and gain, and can even be made to emit beams of more than one wavelength at the same time.

The laser is perhaps the most important product of atomic physics. Traditionally, laser design has relied on careful choice of materials, each with fixed lasing properties, but on page 777 of this issue¹ J. Faist *et al.* describe a much more versatile device, whose properties can be changed at will by applying a small voltage.

In a sense, the laser began more than 80 years ago, when Einstein realized that the decay of an electron in an excited state would be more probable if there were already a photon nearby with an energy matching the decay energy. The electron is stimulated to emit a new photon in phase with the first, so the two light waves reinforce (interfere constructively), and both can cause subsequent stimulated emission from other excited atoms. This is the key process of the laser.

Another condition necessary for decay is that the two electron states be spatially close. The wavefunctions for the states must overlap, because the electron cannot hop great distances to reach the lower-energy state. This is not a problem for an electron decaying to a different state within the same atom, but it is a key question in some commercially important lasers.

Lasing is one matter to imagine and quite another to accomplish. It was not until the late 1950s that the first lasers were built — table-sized instruments based first on emission lines from molecules in vapour, and somewhat later on ruby crystals. Using light or a chemical process to 'pump' the electrons into a long-lived excited state, one creates an overload condition, called a population inversion, which results in an avalanche of electrons falling back into a lower state and emitting a flood of photons. The avalanche is self-triggered, in that the decay of the first electrons stimulates the others to fall and emit their photons.

For the laser to work (to have gain), enough photons must be produced to keep up with the losses from absorption within the device, and with the escape of light into the outside world, where it can be put to practical use (such as reading your CD or correcting the shape of your eye's lens). Usually this is simply a question of choosing the right element and pumping up the population of the excited state enough before the avalanche begins. Enclosing the material in

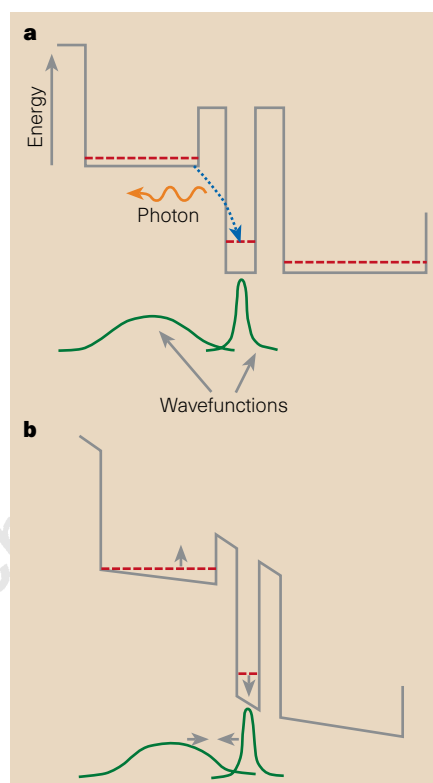


Figure 1 Cascade lasers, old and new. Light is emitted when the electrons fall from the higher-energy, wider semiconductor layer into the lower-energy, narrower layer. The energy difference — which determines wavelength — is the vertical separation of the dashed lines; and the overlap — which determines transition probability and therefore gain — is the extent to which the 'tails' of the wavefunctions span the same region of space. a, In a conventional cascade laser, with no applied voltage, the energy and overlap are fixed. b, In the new device¹, a small applied voltage shears the energy diagram, so both wavelength and gain can be tuned easily.

a leaky cavity then permits the photons to bounce back and forth many times before escaping to the outside, so that a few photons are multiplied into a very large number, all in phase and at the same energy.

The most commonplace form of the laser was invented some years later. In carefully tailored semiconductor crystals, the energy states (bands) of the semiconductors are reservoirs of electrons that can be supplied by electrical currents fed to the device. This

overcomes some of the problem of needing to pump the population of a fortuitous metastable state. More importantly, it permits simple fabrication of small lasers, and that has had considerable commercial repercussions. Again, however, the question of gain is decided by material properties.

The wavelength of the photons is fixed by the energy difference between the excited and low-energy states in any material — one might think of a waterfall whose height is the energy difference. In the quantum cascade laser², different semiconductor layers are laminated together, and electrons decay from a state in one layer to a lower one in the next (Fig. 1a). A series of electron reservoirs are emptied successively downstream, to form many lasers acting as one; there are several identical waterfalls, one after the other. A high-power version of this cascade scheme has now been demonstrated³. The output wavelength can be designed by the choice of a particular set of semiconductors as constituent layers, and for a given choice, the wavelength cannot be adjusted by an appreciable amount.

The new device¹ is a variation of the cascade laser. It too contains several repeated laminations of different semiconductors, AlInAs and InGaAs. In the new laser, however, a voltage applied to the device can be used to change the energy difference between the excited and low-energy states, as well as the amount that the two states overlap; and the spatial overlap determines the transition probability, and therefore the gain. The voltage shears the energy diagram (Fig. 1b) so that the energy difference increases and the states move closer together in space. Instead of simply over-pumping the population to create the avalanche, one can adjust the probability of decay, while tuning the energy difference and the photon wavelength — by more than 5 per cent in this first prototype device. Now one has a series of waterfalls whose heights are adjustable, so, for example, one can change the wavelength of the output light simply by adjusting a quite modest voltage across the device, or even operate two or more parts of the same device independently at different wavelengths simply by applying different voltages to them.

These tunable and polychromatic devices should be useful in many applications, such as spectroscopy and communications. They are poised at the edge of an immense commercial market: one need only calculate the number of telephones, modems and CD-based instruments to foresee great financial and social profits from a new and more versatile laser. □

Sean Washburn is in the Department of Physics and Astronomy, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.

1. Faist, J. *et al.* *Nature* **387**, 777–782 (1997).

2. Faist, J. *et al.* *Science* **264**, 553–556 (1994).

3. Scarmarcio, G. *et al.* *Science* **276**, 773–776 (1997).

Experimental psychology

In a blink of the mind's eye

Jeremy M. Wolfe

Imagine a stream of characters appearing, one at a time, on a computer screen at a rate of about eight per second: P-D-E-X-G-4-F-K. You can read characters at this rate — if asked, you could successfully push a button when you saw a 4. But if you were asked to report on Xs and 4s, you would fail miserably: specifically, you would be unable to respond to the 4 when it appeared, a quarter of a second after the X. Something about the extra attention that is given to the X makes it very hard to respond to other items for the next several hundred milliseconds. Known as the ‘attentional blink’^{1,2}, this is one of the more interesting attentional phenomena to have been described in recent years.

Two papers in this issue offer new insight into what is and is not possible during an attentional blink. Joseph *et al.*³ (page 805) find that attentional blink makes it impossible to perform some very easy visual tasks that were thought not to require attention. In contrast, Duncan *et al.*⁴ (page 808) show that a blink of the mind's eye does not produce a blink of the mind's ear, and vice versa. Both of these results provide new information about attentional blink, but they are more valuable for the broader light that they shed on conscious visual experience (in the paper by Joseph *et al.*), and on the nature of attention (in the case of the paper by Duncan *et al.*).

When we use the term ‘attention’ colloquially — as in “pay attention” — we treat it as a single entity. In fact, whether attention is singular or plural has long been debated. Consider: you are attending to this text. You could, without moving your eyes, attend to some other visual stimulus. You could also attend to the pressure of your posterior on the seat. You were not attending to that tactile stimulus a moment ago — now you are. In the first case, attention is deployed within a single sensory modality whereas in the second, attention is moved between modalities. Does a single, limited attentional resource select prose or posterior as needed? Or, alternatively, are there different pools of attentional resources for different senses?

Duncan *et al.*⁴ provide clear evidence that there are separate resources — at least for vision and hearing. The study is elegantly simple. Two streams of stimuli are presented to the subject, and each can be visual or auditory. One target item appears in each stream, and the subject just reports the targets. The authors found that attentional blink occurs when both streams are visual or when both streams are auditory. However, if one stream is visual and the other is auditory, there is no blink. In other words, when a visual target consumes visual attentional resources, it

leaves resources available for an auditory task. But those resources must be specifically auditory, because they are not available for a second visual task.

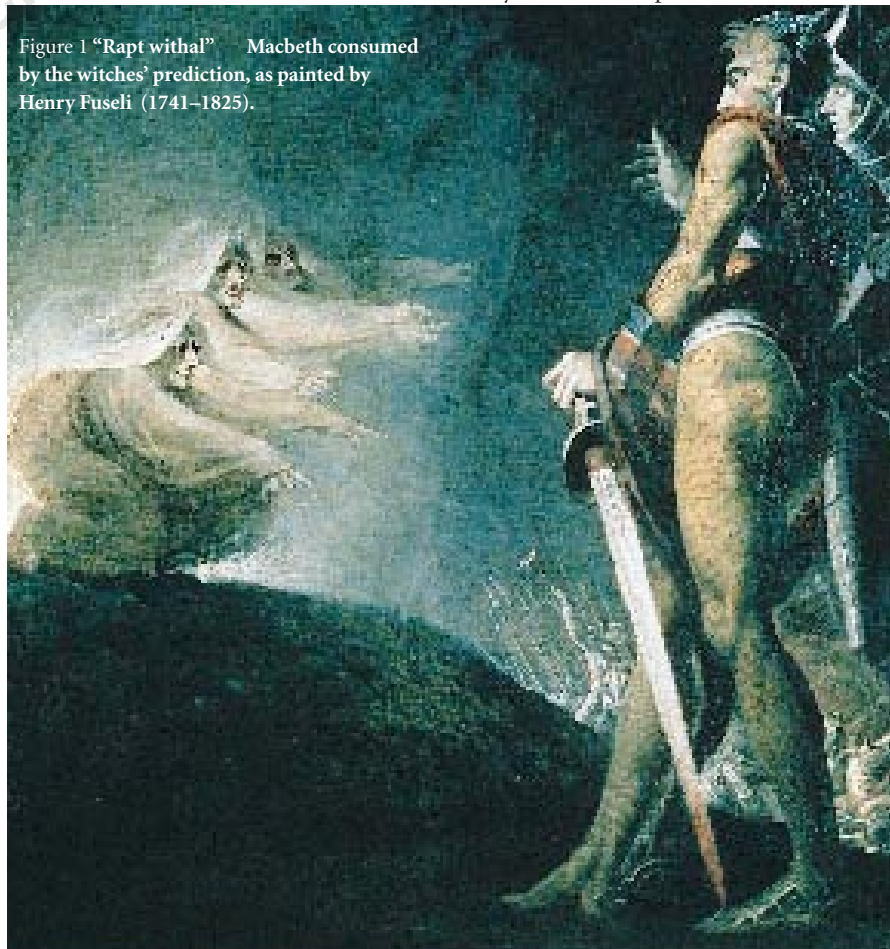
Although this result shows that there are modality-specific attentional resources, it does not rule out the existence of global attentional limitations that can work across modalities. Shakespeare provides an illustration, as depicted in the scene reproduced below, when the witches prophesy that Macbeth shall be “thane of Cawdor ... and ... king hereafter”. Macbeth's attention is so consumed by this “great prediction of noble having and of royal hope, that he seems rapt withal” (Act 1, scene 3). He seems not to hear or to see — some global attentional process has restricted his consciousness to the inner world of his thoughts.

In the other study, Joseph *et al.*³ investigate ‘preattentive vision’. They asked subjects to find a right-tilted target among a number of left-tilted stimuli. In such tasks, the ability to find the target does not depend on the number of left-tilted stimuli. This has been called preattentive vision, because subjects do not need to attend to each item in turn:

the target simply ‘pops out’, as if the entire visual field were processed in parallel^{5,6}. This has sometimes been taken to mean that such tasks require no attention at any stage, from input to motor response, as if there were separate attentive and preattentive visual systems. Indeed, there is reasonable evidence that searches for features such as a tilted target are possible while attention is tied up elsewhere⁷. Joseph *et al.* neatly disproved this by having subjects perform the feature search at the same time as they performed an attentional-blink task (Fig. 2). They found that in trials where the search stimuli appeared during a blink, accuracy for the simple search task was severely impaired.

There are at least two ways to think about this result. Both challenge our usual experience of a visual world, full of visible and meaningful objects. The result could be seen as evidence for what Mack and her colleagues call ‘inattention blindness’⁸. They argue that you can only consciously perceive what is currently attended. Alternatively, the result could be evidence for what I would like to call ‘inattentional amnesia’⁹. Attentional blink may make it impossible to attend or respond to other stimuli for several hundred milliseconds. Those blinked stimuli might be seen, but forgotten by the end of the blink. There is evidence that words presented during attentional blink can be read, even if they cannot be reported after the blink¹⁰.

Figure 1 “Rapt withal” Macbeth consumed by the witches’ prediction, as painted by Henry Fuseli (1741–1825).



NATIONAL TRUST, PETWORTH HOUSE/BRIDGEMAN ART LIBRARY, LONDON

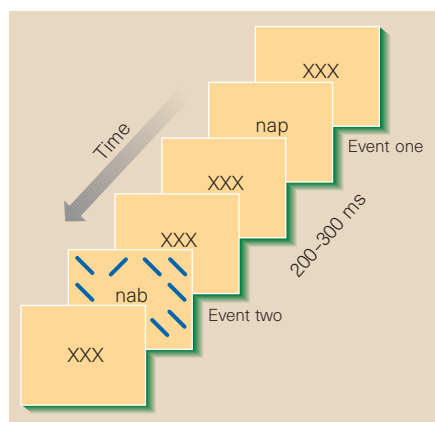


Figure 2 Suppose you were asked to respond to events one and two in a stream of stimuli. Duncan *et al.*⁴ found that subjects missed the second word if both words were visual stimuli or if both words were auditory stimuli, due to the phenomenon known as 'attentional blink'. But subjects could identify the second word if one was auditory and the other was visual. Joseph *et al.*³ found that if event one was visual, subjects could not say whether there was a right-tilted line in the event two display, indicating that attentional blink makes it impossible to carry out easy visual tasks that were previously thought not to require attention.

Moreover, unattended stimuli can produce geometric visual illusions¹¹.

Vivid demonstrations from several labs have pointed to the fleeting nature of visual experience^{12–14}. For instance, Rensink *et al.*¹⁵ presented observers with a scene. After a very brief, 80-millisecond blank period, the same scene was presented in the same location with a single change. Even though observers could clearly perceive and remember the gist of the scene in a second, it took many alternations of the two versions before the change was noticed — even when that change was quite dramatic. For example, subjects failed to readily notice a jet engine appearing and disappearing in a scene where the plane was the most prominent object¹⁵. Although observers may know the gist of the scene, only the present contents of attention can be monitored for change.

In sum, the message of the paper by Joseph *et al.* is that attention is required to carry even the most basic of visual tasks through to completion. If a stimulus is presented while attention is summoned elsewhere, "ere a man hath power to say 'Behold!' The jaws of darkness do devour it up: so quick bright things come to confusion" — although Shakespeare's Lysander is probably not discussing the allocation of visual attention when he speaks these lines in *A Midsummer Night's Dream*. □

Jeremy M. Wolfe is at the Center for Ophthalmic Research, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, Massachusetts 02115, USA.

e-mail: wolfe@search.bwh.harvard.edu

1. Raymond, J. E., Shapiro, K. L. & Arnell, K. M. *J. Exp. Psychol. Hum. Percept. Perform.* **18**, 849–860 (1992).
2. Chun, M. M. & Potter, M. C. *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 109–127 (1995).
3. Joseph, J. S., Chun, M. M. & Nakayama, K. *Nature* **387**, 805–807 (1997).
4. Duncan, J., Martens, S. & Ward, R. *Nature* **387**, 808–810 (1997).
5. Treisman, A. & Gelade, G. *Cogn. Psychol.* **12**, 97–136 (1980).
6. Wolfe, J. M. in *Attention* (ed. Pashler, H.) (Univ. College Press, London, 1997).
7. Braun, J. & Sagi, D. V. *Percept. Psychophys.* **48**, 45–58 (1990).
8. Mack, A., Tang, B., Tuna, R. & Kahn, S. *Cogn. Psychol.* **24**, 475–501 (1992).
9. Wolfe, J. M. in *Fleeting Memories* (ed. Coltheart, V.) (MIT, Cambridge, MA, 1997).
10. Luck, S. J., Vogel, E. K. & Shapiro, K. L. *Nature* **382**, 616–618 (1996).
11. Moore, C. M. & Egeth, H. J. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 339–352 (1997).
12. Grimes, J. in *Perception* (ed. Akins, K.) 89–110 (Oxford Univ. Press, New York, 1996).
13. Simons, D. J. *Psychol. Sci.* **7**, 301–305 (1996).
14. Irwin, D. E. *Curr. Dir. Psychol. Sci.* **5**, 94–100 (1996).
15. Rensink, R., O'Regan, J. K. & Clark, J. J. *Invest. Ophthalm. Vis. Sci.* **37**, S213 (1996).

Gamma-ray astronomy

New sights in the high-energy sky

Hans Bloemen

Gamma-ray astrophysics deals with the world of extremes and fundamentals, from the creation of the elements to the disappearance of matter into black holes. γ -rays are the most energetic electromagnetic radiation, and reveal the most violent phenomena. Six years after the launch of NASA's Compton Gamma Ray Observatory, a 15-ton spacecraft containing the most advanced γ -ray instruments, astrophysicists met* to assess its achievements so far. In some respects, the observations have confirmed theory and expectations, but the γ -ray sky appears to be even richer than anticipated. An unexpected 'fountain' of antimatter far above the Milky Way has been reported, many active galactic nuclei and more than a hundred unidentifiable point sources have been detected, and increasing evidence for γ -ray lines has been seen.

The Compton Observatory carries four instruments, each with its own specific goals and accomplishments. BATSE (Burst and Transient Source Experiment) made us realize that γ -ray bursts are among the most puzzling phenomena in astrophysics. The Energetic Gamma Ray Experiment Telescope (EGRET) and the Compton Telescope (COMPTEL) have mapped for the first time the entire γ -ray sky, from ~ 10 GeV (giga-electron volts) down to the MeV (mega-electron volts) regime, where nuclear decay and nuclear interactions produce γ -ray lines. An intense glow from the Milky Way dominates this whole spectral region, which can largely

be attributed to diffuse emission from cosmic rays hitting gas nuclei and photons in interstellar space. This reveals the distribution of cosmic-ray particles throughout the Galaxy, which is vital information for understanding their origin.

But a growing number of γ -ray sources is being seen, amounting to over 200 now. A small subset can be identified with radio pulsars, not from their positions (which can only be roughly determined by current γ -ray telescopes), but through their distinctive timing signature. And intense γ radiation from more than 50 active galaxies has been discovered, from outflowing plasma jets pointing in our direction, which probably originate from massive black holes (of the order of 100 million solar masses) in the galactic nuclei. Less massive ones in our Galaxy (of a few solar masses or so) are occasionally bright sources in the X-ray and soft γ -ray sky, when a giant extra scoop of ambient matter is accreted. But most of the γ -ray sources remain unidentified. Given their distribution on the sky (loosely following the Milky Way), they are probably within our Galaxy. They are likely to be massive-star remnants of some sort, with pulsars perhaps forming a subclass.

γ -ray spectroscopy is carried out by COMPTEL and by the fourth instrument, OSSE (Oriented Scintillation Spectrometer Experiment). Lines in the γ -ray spectrum are the signatures of nuclear-decay and nuclear-interaction processes. They can provide direct, and often unique, information on many important problems in high-energy astrophysics, including nucleosynthesis,

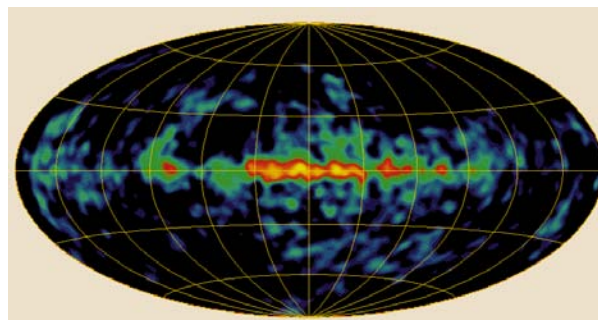


Figure 1 The Milky Way in the light of the 1.809-MeV line from the nuclear decay of ^{26}Al , which has a lifetime of about a million years, obtained with the COMPTEL instrument. This traces sites of recent nucleosynthesis in the Galaxy.

*Fourth Compton Symposium, Williamsburg, Virginia, USA, 28–30 April 1997.

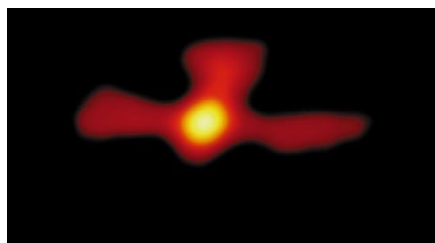


Figure 2 The distribution of positrons near the centre of our Galaxy, obtained with OSSE. The brightest feature is the Galactic nucleus, and the horizontal structure lies along the Milky Way. The newly discovered antimatter cloud, or 'fountain', is above the centre.

particle acceleration, and physical processes near compact objects.

By imaging nuclear-decay γ -rays from radioactive isotopes such as ^{26}Al and ^{44}Ti , we can now see directly where in our Galaxy new elements are being formed and mixed into the interstellar medium by massive stars and supernovae (Fig. 1). Furthermore, we can witness nuclear interactions through nuclear de-excitation lines from, for instance, ^{12}C and ^{16}O . Potential production sites are in violent interstellar environments and near accreting compact objects. Evidence for these spectral features (well known from solar flares) is observed in emission from the nearest site of massive-star formation in giant molecular clouds, the Orion complex, and can now be seen in the large-scale γ -ray glow from the Milky Way as well.

OSSE caused recent excitement in this field. For the past 25 years, more than a dozen balloon and satellite experiments have seen 511-keV emission from the general direction of the Galactic Centre, due to the annihilation of positrons with electrons. Instruments with large fields of view seemed to detect the strongest signal, indicating the presence of more than a single point source. Soon after the launch of the Compton Observatory, OSSE showed that the emission region is indeed extended. An intense glow from the nucleus was seen, with some extension along the Milky Way. But now OSSE has found a new feature (W. R. Purcell, Northwestern Univ.): strong 511-keV emission extends up to 10° above the Galactic Centre (Fig. 2). A 'fountain' of hot gas laden with electron-positron pairs may be rising from the nucleus (C. D. Dermer and J. G. Skibo, Naval Research Lab.), with positrons losing energy and annihilating as they are convected upwards with the gas.

A recent starburst episode near the Galactic Centre could be driving the fountain. And the positrons may come from radioactive material freshly synthesized in stellar explosions associated with the same star-forming activity. The signature of ^{26}Al , for example, might be detectable by future instruments. But nuclear interactions and jets from black holes could also contribute to the positron production.

On the other hand, the enhanced emission may not be related to the nucleus, but instead betray a closer, localized site of positron production and annihilation. If that is the case, other, similar sites may be seen in forthcoming observations.

The Compton Observatory has put γ radiation into the mainstream of astronomy, and it should still solve some of the questions it has raised — after a re-boost into higher orbit was completed on 3 June it is ready for maybe another decade of observations. But more sensitive instruments are needed for radical progress. A new, ten-times-better Compton

Observatory, covering again the low-to-high-energy γ -ray regime, would be colossal, and so is not a realistic option. So a difficult choice must be made: an imaging MeV spectrometer is highly desirable, but so is a telescope to map the high-energy γ -ray sky. As an important next step, the European Space Agency's forthcoming INTEGRAL mission (scheduled for launch in 2001) will drastically improve on spatial and spectral resolution with good sensitivity up to a few MeV. □

Hans Bloemen is at the SRON Space Research Organization of the Netherlands, Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands.

Ecology

Insect pollinators see the light

Peter D. Moore

The complexity of woodland microclimates has long been a source of fascination for ecologists. The patterns of light on a forest floor produce a constantly changing mosaic of intensities that result in marked — and sometimes rapid — changes in temperature and humidity. These factors contribute to a rich and varying assemblage of microhabitats for plants, animals and microbes, and it is easy to understand how such complexity can support a great diversity of species, especially in the stratified architecture of tropical rainforests. But the degree to which microclimatic complexity influences the interactions between organisms and, in the longer term, coevolution, may not be adequately appreciated.

In a recent issue of *Oikos*, Carlos Herrera¹ has examined the forest-floor irradiance mosaics in open oak/pine woodlands in southern Spain. He has found that the patches of sun and shade are important in determining patterns of pollination for ground-dwelling plants and also, perhaps, for the evolutionary development of the plants themselves.

Plant ecologists have recognized the physiological significance of patchiness in woodland light-climates, particularly in determining the intensity and duration of the solar energy that is available for photosynthesis. Sun flecks — direct patches of sunlight that penetrate gaps in the canopy and strike the forest floor — have been examined in relation to the photosynthetic patterns of low-lying leaves. The overall energy that is supplied by sun flecks has been estimated to account for 10–40 per cent of the forest floor's total energy budget^{2,3}. The variation between habitats was thought to be due to such factors as latitude (angle of the Sun), canopy architecture (tree density, height and stratification), wind and cloud cover. But the sudden arrival of a high-intensity patch of light on a plant that is accustomed to life in deep shade can prove damaging, and some

plants, such as the wood sorrells (*Oxalis* spp.), respond by rapidly lowering their leaves into a vertical position.

Insects are also sensitive to variations in woodland microclimate, both in space and time. Patterns of light intensity can be particularly important in the selection of flowers by pollinators⁴. Pollinator choice may, in fact, be more dependent on the microhabitat of the individual plant than on more obvious influences such as the supply of nectar available. Is this simply a matter of energetics? Do some pollinators prefer to work at gathering food in high-energy, sunlit spots, while others prefer low-energy locations with lower temperatures and higher humidity? If so, could this selection be a consequence of the differing thermal biology of the individual pollinator species?

To answer these questions, Herrera¹ has studied the pollination biology of lavender (*Lavandula latifolia*) in the understorey of open woodland of evergreen oak (*Quercus rotundifolia*) and pine (*Pinus nigra*) in southern Spain. Within this patchy habitat, he established a transect of 60 stations at which irradiance was measured over several consecutive days. In the more shaded areas of the woodland understorey, irradiance levels were generally below 200 W m^{-2} and they rarely exceeded $1,000 \text{ W m}^{-2}$. By contrast, the more open sites spent about 25 per cent of the time at light intensities greater than $1,000 \text{ W m}^{-2}$.

Insect pollinators visiting the lavender were netted throughout the day, and their thoracic temperature was recorded within five seconds of capture, using a needle microprobe. Concentrating on the hymenopterans (bees) and the dipterans (flies), Herrera found that there are strong links between species and habitat conditions. Two of the fly species, for example, foraged only in shaded conditions ($<130 \text{ W m}^{-2}$), whereas two of the bee species were found only under conditions of high irradiation ($>800 \text{ W m}^{-2}$). In general, the flies were

restricted to the more shady environments.

Pollinator species varied considerably in their thoracic temperatures: mean values for different species ranged from 27 °C to 43 °C, with the hymenopterans having consistently higher thoracic temperatures than the dipterans. Similarly, the excess temperature of the thorax over the air temperature at the point of capture was greater in the hymenopterans — often greater than 15 °C, compared with less than 10 °C for the dipterans. Herrera also found that the tendency to generate a thoracic temperature that is greater than that of ambient air is positively correlated with body size, because the processes of warming up and cooling down are related to body form and mass. The conclusion is that the sun-seekers are generally smaller and maintain lower thoracic temperatures than the shade-preferring taxa, which are larger and can keep their body temperatures well above their surroundings. So, the thermal biology of insect species determines what micro-habitat they select for their foraging activities. This means that a plant species such as lavender, which can grow and flower both in shaded and open-sun conditions, attracts a different set of pollinators depending on the micro-climate of its habitat: a lavender plant in a sunny spot will be visited mainly by small bees, whereas lavender in the shade will attract more dipterans and larger bees.

The biological and evolutionary implications of these findings are considerable. Various insects will have different seasonal abundances, pollination behaviour, efficiency in pollen collection and delivery, and distances over which they travel between plants. All of these factors will present the host plant with a number of selective pressures to which, over

evolutionary time, it will respond. And because these pressures will differ between 'sun' and 'shade' populations of the plant, pollination could be an important influence on the course of evolutionary development. Previously, the existence of sun and shade ecotypes (or even closely related species) of plants has usually been explained in terms of their photosynthetic physiology, including energy trapping and their ability to maintain a positive carbon balance under different light

climates⁵. But the new study by Herrera¹ shows that botanists cannot afford to neglect the energetic properties and predilections of the insect visitors in the process of speciation. □

Peter D. Moore is in the Division of Life Sciences, King's College, Campden Hill Road, London W8 7AH, UK.

1. Herrera, C. M. *Oikos* **78**, 601–611 (1997).
2. Evans, G. C. *J. Ecol.* **44**, 391–428 (1956).
3. Anderson, M. C. *Biol. Rev.* **39**, 425–485 (1964).
4. Herrera, C. M. *Ecology* **76**, 1516–1524 (1995).
5. Bjorkman, O. *Physiol. Planta* **21**, 84–89 (1968).

Demography

Taking the measure of uncertainty

Shripad Tuljapurkar

Judged on performance, population forecasts (usually published by institutions such as the United Nations or census bureaux) have a mixed record. They are no worse than forecasts by economists, meteorologists and others who have to deal with complex and partially understood systems, but demographers have been understandably concerned about error in their predictions. Part of their response is a new wave of demographic work focused on the forecasting of uncertainty, *per se*. Such work aims to forecast a range of demographic outcomes along with associated probabilities, rather than one prediction that will almost surely be wrong.

This new approach is illustrated by Lutz, Sanderson and Scherbov (page 803 of this issue¹), who present a forecast of world population through the year 2050 and up to 2100. They conclude that the odds of world population doubling by 2050 are less than a third, whereas a doubling of the fraction of the population that is over the age of 60 is essentially a sure thing. Their results stem from a consensus of expert opinion that human fertility will continue to fall everywhere, trailing the decline of mortality by about a half-century. Their result is also noteworthy for its assignment of explicit probabilities to alternative futures.

Traditional forecasts use a 'high–medium–low' method, making a central forecast bracketed by two variants. Although the range between 'high' and 'low' indicates uncertainty, no probabilities are associated with the alternative outcomes, so it is difficult to interpret, employ and evaluate them². In contrast, a probabilistic forecast is valuable to anyone who must make a decision that turns on future events, because one may compute the odds of happy and nasty outcomes, and turn decisions into informed gambles. This is no radical idea; investors, for example, routinely use estimated uncertainty to shape their decisions³. But the idea is new to demography, for there are technical challenges, and the forecaster has to educate the people who actually use the forecasts.

A forecaster begins with today's known

conditions and projects to a future date that is characterized by unknown demographic conditions such as the level of fertility or mortality. The expert makes assertions about demographic characteristics (for example the level of fertility or the expected length of life) at the end-point, and the trajectories along which those characteristics will change between start and end. I refer to the first of these as 'static' scenarios, because they describe possible conditions at a single future time, and to the second as 'dynamic' scenarios. Finally, one must attach *a priori* probabilities to each static or dynamic scenario.

Lutz *et al.* establish static scenarios, alternative conditions of fertility and mortality in 2050, with attached probabilities. To make a forecast, select one final fertility and mortality ('there'), specify a time-course along which the starting fertility and mortality ('here') changes to 'there', and project the starting population over that time-course. Repeat for each ending scenario, and weight the projected population by the probability associated with the ending scenario. Astute readers will ask how probabilities are assigned to scenarios, and how one gets to each 'there' from 'here'? Lutz *et al.* employ subjective probabilities based on a sampling of expert opinion; this is a version of what used to be called (with some chutzpah) the Delphi method. A more systematic approach uses a retrospective analysis of expert predictions to quantify their bias and error^{4,5}, and is more objective about assigning probabilities.

The dynamics of getting 'there' from 'here' are largely ignored in the static approach (it is usual to pick some smooth trajectory between initial and final conditions). The usual reasons are a lack of information or a belief that it isn't really important. Neither is convincing in a probabilistic forecast that is motivated precisely by uncertainty.

In contrast, dynamic forecasts employ a stochastic model of changes in fertility, mortality and migration, typically fitted to historical data. The resulting models have uncertainty embedded within them, as reflected in historical change, and yield



Figure 1 When sunny gets blue — plants such as lavender, which can grow and flower in both sunny and shaded areas, are visited by different insect pollinators depending on light intensity. According to Herrera¹, whereas large insects such as flies visit plants in shaded areas, smaller insects preferentially visit plants in sunny spots.

probabilities for alternative trajectories of future population^{6,7}. The trajectories in such models typically contain brief episodes of rapid change and are far from smooth. Such an approach is particularly valuable in assessing social programmes (for instance the US social security system) whose future is sensitive to the timing and pattern of demographic change.

Dynamic approaches tend to be 'formal', requiring an explicit mapping from historical information into processes of change. Static scenarios tend to be 'subjective', because they rarely make explicit how history is mapped into the future. Dynamic approaches are criticized for an over-reliance on history, on the grounds that the future may hold unknown revolutionary change. That may be true, but the history of this century is surely characterized by tumult and sudden change. Besides, the static approach relies just as heavily on analysing history — which is what makes experts.

The difference between static and dynamic approaches is partly a matter of timescales: fertility change between 1997 and 2050, for instance, is driven by processes that work over two or more human generations. The actual course of transition is marked by changes on a generational timescale, of the sort whose signature is writ large in the demographic history of any country. The overarching goal is the enterprise of developing probabilistic forecasts that are directly useful tools for decision and analy-

sis. With luck, demographers will point the way for ecological and environmental forecasts, which are also bedevilled by uncertainty.

Lutz *et al.* make a persuasive case that global population growth is slowing down, which should soothe those who see population as the primary cause of the world's ills. But they predict high odds of at least a half-century of substantial global population growth as well as a tremendous increase in the proportion of older people in every country. Those involved in economic, social and population policy still have lots to worry about. International economic policy has a large part to play in helping today's most populous countries stay on their courses of fertility decline. This is an important role if we are to steer away from the unlikely but real possibility that fertility will start to rise again. For individuals, families and countries everywhere, the largest question of the next few decades will almost surely be, how to age gracefully. □

Shripad Tuljapurkar is at Mountain View Research, 2251 Grant Road, Suite A, Los Altos, California 94024, USA.

e-mail: tulja@mvr.org

1. Lutz, W. L., Sanderson, W. & Scherbov, S. *Nature* **387**, 803–805 (1997).
2. Lee, R. D. *Int. J. Forecast.* **8**, 315–327 (1992).
3. Malkiel, B. G. *A Random Walk Down Wall Street* 3rd edn (Norton, New York, 1992).
4. Keilman, N. *Uncertainty in Population Forecasting* (Swets & Zeitlinger, Amsterdam, 1990).
5. Alho, J. A. *J. R. Statist. Soc. A* **80**, 306–314 (1996).
6. Lee, R. D. & Tuljapurkar, S. *J. Am. Statist. Ass.* **89**, 1175–1189 (1994).
7. Alho, J. A. & Spencer, B. J. *Official Stats* **7**, 295–310 (1991).

Lubricants

Super slippery solids

Somuri Prasad and Jeffrey Zabinski

It would appear that primitive man had a clear understanding of the importance of friction and lubrication — transportation of massive objects by the Egyptians in 2000 BC, for example, was facilitated by using sledges dragged over lubricated wooden boards. The role of a lubricant is similar to that of a peace-keeping force: its main function is to prevent the opposing surfaces from coming into close contact with each other at the atomic level. Most lubricants are liquids, but when the operating conditions (such as high or low temperature, or vacuum) are beyond the liquid realm, attention turns to solid materials. The most common solid lubricants are graphite and the transition-metal dichalcogenides, MX_2 (where M is molybdenum or tungsten, and X is sulphur, selenium or tellurium). But both graphite and the metal dichalcogenides have certain inherent deficiencies. The synthesis of hollow nanoparticles of tungsten disulphide (WS_2) similar to fullerenes and nanotubes, reported by Tenne and co-workers on page 791 of this issue¹, offers some exciting possibilities for a new generation of solid lubricants.

First, how do traditional solid lubricants work? The low friction of both graphite and metal dichalcogenides is usually due to interplanar mechanical weakness, intrinsic to their crystal structures. For example, WS_2 crystallizes in the hexagonal structure, in which a sheet of tungsten atoms is sandwiched between two hexagonally packed sulphur layers (Fig. 1, overleaf). The bonding within the S–W–S sandwich is covalent, whereas the sandwiches themselves are held together by weak Van der Waals forces, resulting in interplanar mechanical weakness. Under the action of a shear force, intracrystalline slip occurs in the weak interplanar regions. This mechanism is responsible for the formation of smooth transfer films by wear: the new surfaces, created by separating the weakly bonded sandwiches, are quite inert. They can easily slide back and forth over one another (by intercrystalline slip), thereby providing lubrication.

But there is a major obstacle to lubrication by metal dichalcogenides: the presence of unsaturated or dangling bonds. In a typical layered structure ($2H-WS_2$), the obvious



100 YEARS AGO

Another name must be added to the long list of martyrs who have given up their lives while endeavouring to effect the conquest of the air. The latest victim is Dr. Wölfert, who had devoted many years to the problem of aerial navigation, and who claimed to have invented a navigable balloon. The Berlin correspondent of the *Times* says that Dr. Wölfert had made an arrangement with the officers of the ballooning section of the army to put his invention to a practical test at Tempelhof on Saturday last. ... At first the balloon ascended steadily and began to make good progress... Suddenly, however, when the balloon was sailing at a height of about 1000 feet, flames shot up from the car and the balloon exploded with a loud report and was precipitated, a burning mass, into a wood-yard below. The mounted officers hurried to the spot, and, after the flames had with great difficulty been partially extinguished, the mutilated remains of Dr. Wölfert and his companion were found amidst the ruins of the car. It is believed that the valve of the balloon was opened with the intention of descending, and that the gas, in escaping from the balloon, became ignited by the benzene. From *Nature* 17 June 1897.

50 YEARS AGO

The Committee's programme, as outlined in communications recently received from Prof. Einstein, is to see that the following simple facts, which are accepted by all scientific workers, are given the widest possible publicity. They are: (1) that atomic bombs can now be made cheaply and in quantity, with future bombs likely to be even more powerful and destructive; (2) that there is no military defence against atomic bombs; (3) that other nations can discover for themselves the processes kept secret by the United States; (4) that preparedness against atomic warfare is futile and, if attempted, will ruin the structure of social order; (5) that if war does occur, atomic bombs will definitely be used, and will surely destroy our civilization; and (6) that international control of atomic energy and, ultimately, the elimination of war, is the only solution to the problem. — 'Emergency Committee of Atomic Scientists'

From *Nature* 21 June 1947.

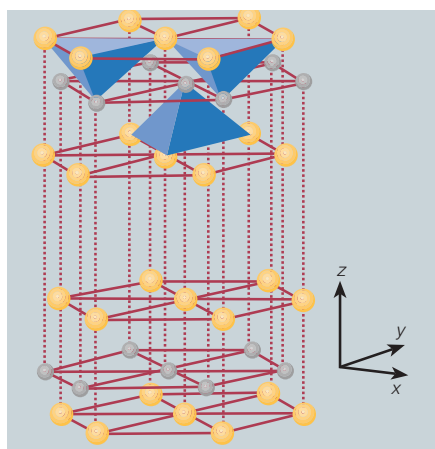


Figure 1 Crystal structure of hexagonal tungsten disulphide, 2H-WS₂.

source of dangling bonds is an edge plane. Furthermore, the propagation of a pre-existing crack in a WS₂ film will break a few covalent bonds and create more unsaturated bonds. If sliding takes place in humid air, such activated surfaces can instantly react with moisture and oxygen in the surrounding environment, forming WO₃. Thus, the tribological (friction and wear) behaviour of metal dichalcogenides is strongly dependent on their environment. For instance, when the test environment is switched from dry nitrogen to humid air², the friction coefficient of a typical WS₂ film can rise from 0.03–0.04 to 0.15–0.20, decreasing its wear life by several orders of magnitude. In contrast, graphite loses its lubricating property in a vacuum, as the complete absence of adsorbed vapours makes it difficult to shear the layers. According to our current understanding, no material can act as a solid lubricant in all these environments. Has such a material now been discovered?

Tenne and co-workers³ first observed closed structures of WS₂, ranging in size from 10 to 100 nm, when they annealed tungsten films at 1,000 °C in a reducing H₂S atmosphere (see Fig. 2). The distance between two fringes in their lattice images is 0.615 nm, the same as the distance between two neighbouring sandwiches in an open 2H-WS₂ structure. Tenne and co-workers³

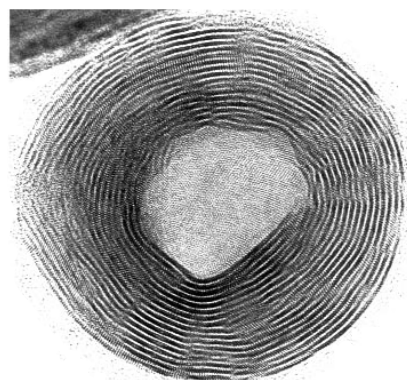


Figure 2 High-resolution electron micrograph of a WS₂ nanoparticle.

argue that the curling and closure of WS₂ networks was probably induced by thermal stresses (1,000 °C is a fairly high temperature) and facilitated by structural defects such as edge dislocations, or by contaminant atoms, such as oxygen, that came from the quartz substrate. The authors have also made nanoparticles of WS₂ (HN-WS₂) in gram quantities by a sol–gel reaction¹.

The dangling bonds that are responsible for causing oxidation in layered 2H-WS₂ structures are no longer present in HN-WS₂ particles. Does this mean that we now have a super solid that can effectively lubricate in all environments? The intracrystalline slip that is responsible for creating smooth transfer films is no longer present in this powder, as appreciable shear between inner and outer closed layers is almost impossible. By the same token, HN-WS₂ powder cannot be burnished like the traditional layered structures. Indeed, Tenne and colleagues¹ had to roughen the substrate and use a drop of mineral oil to avoid runaway of the nanopowder during friction measurements. Yet the friction coefficient and wear rate for nano WS₂ powder measured in oil were superior to those of traditional WS₂ and MoS₂ powders¹. This suggests that a different mechanism is controlling the tribological behaviour of nanoparticles.

The mechanism of lubrication in the case of hollow nanopowder appears to be one of rolling and not sliding. By virtue of their extremely small size, the HN particles can fill the submicrometre-sized valleys between asperities and prevent asperity contact between mating surfaces. This, coupled with their rounded shape, opens the possibility that they undergo rolling friction at the nanometre level. Furthermore, elimination of dangling bonds makes the HN particles chemically inert, so the particles are less likely to stick to the substrate, or to one another. Their hollow-cage structure imparts high elasticity, and elastic deformations (as opposed to inelastic deformations) diminish the energy dissipation associated with friction. These are some of the characteristics that make the hollow nanoparticles of metal dichalcogenides potential candidates for miniature rolling-element bearings.

Perhaps, in the coming decades, we might see a micrometre-sized rolling-element bearing using hollow nanotubes of metal dichalcogenide. But even in the near future, these small, round, inert nanoparticles should find some applications in powder lubrication, and as additives in lubricating oils. □

Somuri Prasad and Jeffrey Zabinski are in the Materials Directorate of the Wright Laboratory, Wright Patterson Air Force Base, Ohio 45433-7750, USA.

1. Rapoport, L. *et al.* *Nature* **387**, 791–793 (1997).

2. Prasad, S. V. & Zabinski, J. S. *J. Mater. Sci. Lett.* **11**, 1413–1415 (1993).

3. Tenne, R. *et al.* *Nature* **360**, 444–446 (1992).

Daedalus

A science futures market

When a manager starts to depend on performance indicators, he is losing his grip. He should know what his concern is up to, without juggling arcane ratios. And once performance indicators are taken seriously, doom is certain. The organization switches its goals from whatever it should be doing, to maximizing the indicators.

On this criterion, scientific research is in a bad way. Performance indicators abound — impact factors, citation counts, publications per head or per grant dollar, and so on. Mutual backscratching cliques cite each other furiously in all their papers; journals parasitized by such publications proudly proclaim their impact factors; grant money flows to individuals and organizations who can show the best performance indicators. Hence the delight and outrage over the recent research-assessment ranking of UK universities.

Daedalus now has a better way. Science is no longer a collective endeavour to understand the physical world; it is a competitive struggle to acquire grant money. The time has come to put it on a sound capitalistic basis. So Daedalus proposes a scientific stock exchange. University departments, and even individual scientists, could issue shares in their research programmes, and shareholders would receive a proportion of each grant subsequently won by the issuer.

At once the 'invisible hand' of the market would work its benevolent magic. Researchers would issue prospectuses arguing the strength of their teams, the brilliance of their hypotheses and the convincing nature of their pilot experiments. The grant-worthiness of such claims would be judged by shrewd and critical scientific investors. The intellectual debate would be sharpened by the attempts of competing shareholders to talk up the projects they had backed, or to criticize ones in which they had taken a 'stag' position. In the same way, the scientific grapevine would instantly evaluate alleged breakthroughs or setbacks, and signal its opinion through the relevant share prices.

Research would be transformed. Any scientist could obtain prompt 'seed-corn money' for his ideas by issuing shares. Progress on all research fronts could be assessed daily in the financial papers. And a whole new scientific career would open up. Elderly researchers would no longer see administration yawning inescapably in front of them. Their experience and insight would gain them steady money punting on the scientific market.

David Jones

Obituary

Keith Roberts Porter (1912–97)

Pioneer in cell biology

Major achievements in science are almost always associated with the names of exceptional scientists. Keith Porter is one such name. Porter, who died on 2 May at a retirement home in Bryn Mawr, Pennsylvania, aged 84, was one of the founding fathers of modern cell biology. History will ascribe a string of discoveries, technical advances and thought-provoking ideas to him.

A native of Nova Scotia, Porter graduated from Acadia University in 1934 and received his doctorate in biology at Harvard University. His first and perhaps most fruitful scientific period started in the early 1940s at the Rockefeller Institute (now Rockefeller University) in New York. While he was studying cells in tissue culture, it occurred to him that they might be spread thinly enough to be observed with a then-revolutionary type of instrument that used electrons instead of visible light for image formation. The boldness of this idea can hardly be imagined today when electron microscopy has become routine. Then, however, it was so radically new that procedures for specimen preparation were non-existent.

So how did one go about preparing a cell for viewing in this alien but powerful new instrument? A good start was to use an approach analogous to light microscopy, only replacing the glass slide with a thin film of plastic and using heavy metal atoms instead of dyes to enhance contrast. The plastic film with cells attached could be picked up on a small wire mesh, and introduced into the electron beam in the hope that some of the cells would come to lie in a hole of the mesh. In this first experiment, one of the cells did, and as Porter recalled “It became the most photographed cell in history” (see the cover of the journal *Molecular Biology of the Cell*, April 1997). This was the birth of whole-mount electron microscopy, a technique that Porter would return to some 30 years later; and it was the first of his many contributions to the development of techniques for biological electron microscopy, the most significant being the construction, together with Joseph Blum, of the first reliable microtome for cutting thin sections.

Back in the 1940s, the fine structure of cells was entirely uncharted territory. The first glimpse of the intricate nature of the internal organization of a cell must have



been exhilarating, comparable to the first look at the night sky with a telescope 300 years earlier. The work of Porter and his colleagues (Albert Claude and Ernest Fullam) paved the way for the concept of compartmentalization of the cell by membranes — a principle so fundamental to biology that it is high-school material today.

From the early 1950s, Porter and George Palade, his congenial colleague and co-leader of the laboratory at the Rockefeller Institute, and a future Nobel laureate, attracted students, postdocs and visiting fellows from all over the world. Porter was an exceptional speaker, generating excitement in first-year students and established scientists alike, and he created an atmosphere where the scientific imagination could thrive. His group made a series of outstanding advances, including the first ultrastructural description of virus particles (1948, with Helen Thompson), the imaging of the 9+2 pattern of cilia and flagella (1954, with Don Fawcett), the description of the sarcoplasmic reticulum of muscle cells (1953, with Stanley Bennett), the first study of fibres in the extracellular matrix (1947, with Bud Hawn), and the recognition that microtubules are ubiquitous structures in the cytoplasm (1963, with Myron Ledbetter). It is amazing how deeply the science of cell biology is rooted in the discoveries of Porter and his talented associates of these early years.

In 1961, at the height of his success at the Rockefeller, Porter left to take the chair of the biology department at Harvard. Only eight years later he moved

on, accepting the challenge of setting up a new biology department at the University of Colorado at Boulder, a task sweetened by the promise of a brand-new building literally constructed around a novel piece of equipment, a two-storey-tall, high-voltage transmission electron microscope.

Once again this new approach — observing whole cells and thick slices of tissue in three dimensions — provided stunning insights into the complexity of cellular organization. Porter became fascinated, even obsessed, with the cytoplasmic ground substance, the *gelée vivante*, between the membrane systems he had first seen a quarter of a century earlier. He coined the term ‘microtrabecular lattice’ to label what he believed was an all-pervading and all-controlling network of fine fibres that organizes and directs many cellular activities. The concept as formulated by Porter was not confirmed, but his vigorous advocacy stimulated enquiry into the nature of the cytoplasmic matrix that resulted in a better understanding of its intricate structure. For example, his idea of the cytoplasm as a functional unit organized, in large part, by the microtubule system has since surfaced in many guises.

Porter the scientist stands out prominently, but Porter the science politician has left his mark as well. For example, he was one of the founders of the American Society for Cell Biology and the *spiritus rector* behind the creation, in 1955, of the *Journal of Cell Biology*, then called the *Journal of Biophysical and Biochemical Cytology*, and now one of the leading journals in the field. Throughout his career, he also convinced funding agencies of the importance of structural analysis for cell biology.

Not surprisingly, as someone who accomplished so much during his 50-year scientific career, Porter also received numerous awards and prizes — except one. Indeed, many of us believe that he should have been among those who received a Nobel prize for contributions to cell biology. Although this honour was denied him, he is assured of the still-greater honour of being remembered through his work. “If I have seen further it is by standing on the shoulders of giants,” wrote Isaac Newton to Robert Hooke. Porter, though small of stature, was one of the giants of cell biology who allows his successors to gain deeper insights into the nature of the cell.

Manfred Schliwa

Manfred Schliwa is in the Adolf-Butenandt-Institut, Zellbiologie, University of Munich, 80336 Munich, Germany.

What keeps sandcastles standing?

Any child playing on the beach knows that the physical properties of wet and dry sand are very different. Wet sand can be used to build sharp-featured sandcastles that would be unstable in dry sand. We have now quantified the effect of adding small quantities of liquid to a granular medium. Nanometre-scale layers of liquid on millimetre-scale grains dramatically increase the repose angle (the steepest stable slope that the substance can form) and allow the development of long-range correlations, or clumps.

Moisture-induced changes in granular media are primarily caused by adhesive forces associated with interstitial liquid bridges between grains. Such effects are significant in industries as diverse as pharmaceuticals, construction and agriculture. Liquid-induced forces also affect experimental studies of the physics of granular media^{1–4}. Previous studies of moisture in granular media have examined only relatively large quantities of water added to highly irregular, porous or water-soluble materials such as coal^{5,6}, sugar⁷, seeds or rock chips^{8,9}. The resulting data are difficult to analyse and do not aid understanding of the physics of wetted granular media.

We have studied the effects of the addition of small quantities of corn oil and vacuum-pump oil (which both have low vapour pressures) to spherical polystyrene beads, which are insoluble in these oils. We measured the angle of repose by the draining-crater method¹⁰, with varying draining apertures and liquid content. Our maximum liquid content, (40 times less than the minimum moisture content of previous studies) corresponds to a liquid coating thickness (t_{liq}) of less than 50 nm on the spheres. We found an enormous increase of the angle of repose (Θ_R) with t_{liq} (Fig. 1a). One surprising feature is that Θ_R seems to increase linearly with t_{liq} up to the point where clumping occurs. The results are independent of oil type and aperture (except for the largest t_{liq}), and Θ_R at $t_{\text{liq}}=0$ agrees with previous measurements by this method¹⁰.

In dry sand, Θ_R is determined by the shape of the grains and by the friction forces. In wet sand, the attractive forces due to interstitial liquid bridges increase the stability of the

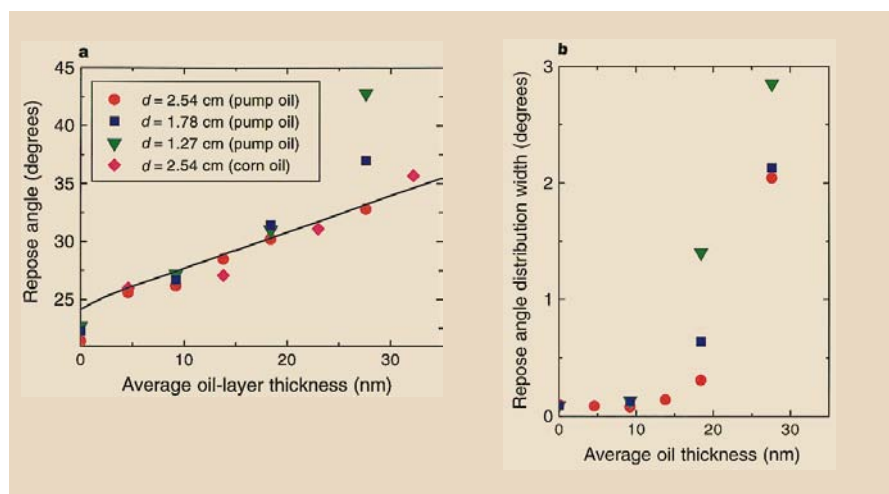


Figure 1a, The angle of repose (Θ_R) as a function of the average liquid-layer thickness on spheres (750 g samples, sphere diameter 0.8 ± 0.2 mm). Draining apertures, d , are listed. Data have been corrected for changes in relative humidity (40–50%) which had a small ($\sim 1^\circ$) effect on Θ_R . Experimental uncertainty is thus roughly 1° . Solid curve, one-parameter theoretical fit. Deviation from the theoretical values for small apertures and large t_{liq} is due to clumping. **b**, The distribution width of measured Θ_R as a function of pump-oil layer thickness. Width corresponds to variations in roughness of the crater surface, which increases with cluster size, as seen with larger t_{liq} .

surface particles and hence Θ_R . Taking into account the observed surface roughness ($\sim 1 \mu\text{m}$) of the spheres, we calculated the inter-particle adhesive force as a function of liquid content¹¹. We then calculated the angle of repose for the wet material using a method based on the stability of the surface particles¹². This model fits the data with a single free parameter corresponding to the volume of liquid in a typical bridge. The fit indicates that the typical volume of a liquid bridge is $\sim 3 \times 10^{-17} \text{ m}^3$ for our maximum t_{liq} , implying that 99.9% of the liquid does not contribute to the adhesive force, possibly because of the surface roughness.

As liquid was added to the spheres, correlated particle clusters (clumps) formed, the size of which increased with liquid content. The presence of such clusters leads to the appearance of an aperture dependence in Θ_R for the largest values of t_{liq} . The development of clustering is also shown by increases in the width of the distribution of values of Θ_R with increasing t_{liq} (Fig. 1b). The width corresponds to variations in the roughness of the crater surface (for $t_{\text{liq}} \geq 20$ nm the craters were noticeably asymmetrical whereas for smaller t_{liq} the surfaces were essentially conical with roughness equivalent to a few sphere diameters).

The development of such clusters appears to be rather sudden, suggesting a transition from a situation where the bulk properties are associated with the dynamics of individual grains to a situation where long-range correlations dominate. Our apparatus entirely failed to drain for larger

t_{liq} (≥ 40 nm) when the size of the clusters approached the aperture size.

Small quantities of wetting liquid can thus dramatically change the properties of granular media, leading to a large increase in the repose angle, clustering and correlation in grain motion. Our results indicate that interstitial liquids can alter many aspects of pattern formation, self-organization^{1–4} and segregation¹³ in granular materials, potentially leading to new physical phenomena not encountered in dry matter.

**D. J. Hornbaker, R. Albert, I. Albert
A.-L. Barabási, P. Schiffer**

Department of Physics,
University of Notre Dame, Notre Dame,
Indiana 46556, USA
e-mail: peter.schiffer.1@nd.edu

- Jaeger, H. M., Lui, C.-H. & Nagel, S. R. *Phys. Rev. Lett.* **62**, 40–43 (1989).
- Alonso, J. J. & Herrmann, H. J. *Phys. Rev. Lett.* **76**, 4911–4914 (1996).
- Umbanhowar, P. B., Melo, F. & Swinney, H. L. *Nature* **382**, 793–796 (1996).
- Jaeger, H. M., Nagel, S. R. & Behringer, R. P. *Rev. Mod. Phys.* **68**, 1259–1273 (1996).
- Standish, N., Yu, A. B. & He, Q. L. *Powder Technol.* **68**, 187–193 (1991).
- Wolf, E. F. & von Hohenleiten, H. L. *Trans. Am. Soc. Mech. Eng.* **67**, 585–599 (1945).
- Craik, D. J. & Miller, B. F. *J. Pharm. Pharmacol.* **10**, 136T–144T (1958).
- Fowler, R. T. & Wyatt, F. A. *Aus. J. Chem. Eng.* **1**, 5–8 (1960).
- Pilpel, N. *Manufact. Chem. Aerosol News* **41**, 19–22 (1970).
- Brown, R. L. & Richards, J. C. *Principles of Powder Mechanics* (Pergamon, Oxford, 1970).
- Eremenko, V. N., Nadich, Yu. V. & Lavrimenko, I. A. *Liquid-Phase Sintering* (Consultants Bureau, New York, 1970).
- Albert, R., Albert, I., Hornbaker, D. J., Schiffer, P. & Barabási, A.-L. *Phys. Rev. Lett.* (submitted).
- Makse, H. A., Havlin, S., King, P. R. & Stanley, H. E. *Nature* **386**, 379–382 (1997).

Natural isotope markers in salmon

To optimize conservation and restoration strategies for fish stocks it is important to identify the rearing habitats that produce the most successful individuals. However, tracking millions of migratory fish through multiple life stages seems to be impossible using conventional techniques. Using differences in the ratio of stable isotopes of strontium (Sr), found naturally in stream water, we have been able to distinguish juvenile Atlantic salmon (*Salmo salar*) from eight of ten rearing sites studied in Vermont streams. This demonstrates the potential for using environmental signals to determine the rearing stream from which salmon originate, and should help to guide restoration activities.

Seven million Atlantic salmon fry (about 30 mm long) are stocked annually into tributaries of the Connecticut River in the United States. Virtually all adults in this restoration effort are captured before returning to their rearing streams and are used in breeding programmes. Therefore, there is currently no way to determine

which rearing stream produced a returning adult salmon. Because the fry are too small to be physically tagged before stocking, we investigated the use of stable isotopes of Sr, which seem to provide an environmental signature¹, to distinguish salmon from different rearing streams.

The isotopic composition of Sr (⁸⁷Sr/⁸⁶Sr, expressed as δ⁸⁷Sr; Table 1) dissolved in stream water reflects that of the watershed geology²⁻⁴. Given the geological diversity across our study region, we expected Sr isotope ratios to vary significantly among tributaries. Within a site, stream-water δ⁸⁷Sr remains relatively constant across seasons and years^{5,6}. Fish take up Sr isotopes in direct proportion to their availability in water, and substitute Sr for calcium in calcified tissues during growth⁷. Because the Sr isotope ratio does not change during uptake and assimilation, δ⁸⁷Sr of fish bones and otoliths (ear-stones) should be the same as that of the stream water⁸.

Earlier attempts to identify natural fish markers have focused on differences in elemental concentrations, for example, distinguishing migratory from non-migratory fish populations using the large difference in Sr concentrations between marine and fresh water^{9,10}. Isotope ratios provide a

much more sensitive method than elemental concentrations for distinguishing multiple populations¹¹.

We collected fish and water from 10 salmon stocking sites in two major watersheds of the Connecticut River. Three months after stocking, the vertebral δ⁸⁷Sr signature in the salmon was the same as the signature of the stream where they were collected. During this time the salmon undergo explosive growth, and the Sr signal of hatchery fry (~4‰) is lost because of exchange and accumulation of Sr from the rearing tributary. We found only minor seasonal fluctuations in stream-water δ⁸⁷Sr values (<0.50‰; Table 1).

The δ⁸⁷Sr values in salmon aged 0 and 1 years from the same sites did not differ, so we pooled all fish from each site and found a strong correlation between the δ⁸⁷Sr value in the stream water and that in the tissues of 90% of fish. We believe that the most likely explanation for the two 'mismatched' fish is that they migrated from an isotopically distinct site. Thus, our results suggest that few juvenile salmon (<10%) move after stocking, but that they may move great distances, as the habitat in which one fish was found was 5 km from the nearest habitat that could have contributed its distinct δ⁸⁷Sr signal.

On the basis of δ⁸⁷Sr values in salmon backbones and water, we separated fish from 10 rearing sites into 8 distinct groups (Table 1). There was significant overlap of δ⁸⁷Sr between only two pairs of sites. For one pair (two sites separated by less than 3 km), the upper site drains into the lower site. The second pair of sites lie further apart but seem to drain geologically similar catchments.

Finally, the δ⁸⁷Sr values of sagittal otoliths closely matched the vertebral signature (Table 1). Otoliths grow in discrete annuli which can be isolated and used to trace the age-specific chemical history of a fish¹². In returning adult salmon, ocean-derived Sr will be in the outermost otolith bands. Because the world's oceans have no measurable spatial or temporal variability in δ⁸⁷Sr, ocean-derived Sr in returning adult salmon can be reliably and easily separated from the freshwater signal. Thus we should be able to distinguish between stocks on the basis of their freshwater signature and to estimate tributary-specific contributions to smolt and adult salmon production.

Brian P. Kennedy*

Carol L. Folt*

Joel D. Blum†

C. Page Chamberlain†

*Department of Biological Sciences, and

†Department of Earth Sciences,

Dartmouth College, Hanover,

New Hampshire 03755, USA

e-mail: bpk@dartmouth.edu

Table 1 Strontium isotope compositions in West and White River drainage basins

Sample site	Sample date	δ ⁸⁷ Sr in water	δ ⁸⁷ Sr in vertebrae	δ ⁸⁷ Sr in otolith
Bethel hatchery	30 May 1995		4.15 ± 0.07	
White River tributaries				
1. First branch ^a	5 May 1995	2.18 ± 0.04		
	13 Sep. 1995	2.06 ± 0.05		
	23 Aug. 1996		1.91 ± 0.05	
2. White River – upper ^b	20 Aug. 1996		12.37 ± 0.05	
3. Third branch ^c	10 Sep. 1996		14.43 ± 0.04	
4. Tweed branch ^d	12 Oct. 1995	19.13 ± 0.04	18.11 ± 0.04	
	12 Sep. 1996		17.34 ± 0.04	
5. Bingo brook ^e	5 May 1995	19.47 ± 0.05		
	24 Aug. 1995	19.97 ± 0.03	19.75 ± 0.04	
	12 Oct. 1995	19.67 ± 0.04	19.79 ± 0.04	
	27 Aug. 1996		18.45 ± 0.02	
	27 Aug. 1996		19.49 ± 0.04	
6. West branch ^e	24 Aug. 1995	20.30 ± 0.04	20.17 ± 0.04	
	12 Oct. 1995	20.19 ± 0.06	19.83 ± 0.04	
	27 Aug. 1996		19.71 ± 0.06	
	27 Aug. 1996		15.76 ± 0.03*	
West River tributaries				
7. Utley brook ^c	19 Oct. 1995	14.96 ± 0.05	14.46 ± 0.05	
	19 Oct. 1995		14.62 ± 0.05	14.20 ± 0.02
8. Baker brook ^f	10 May 1995	22.95 ± 0.06		
9. West River – upper ^d	19 Oct. 1995	24.80 ± 0.59	26.00 ± 0.03	
	19 Oct. 1995		31.93 ± 0.06*	30.78 ± 0.02
10. Marlboro branch ^h	19 Oct. 1995	27.69 ± 0.04	27.58 ± 0.06	
	19 Oct. 1995		27.40 ± 0.15	

Isotope ratios for Sr presented as δ values. δ = ((R_{sample}/R_{standard}) - 1) × 1,000, where R = ⁸⁷Sr/⁸⁶Sr. The Sr isotope reference standard is modern ocean water (R_{standard} = 0.70918) (ref. 13). Reported uncertainties are 2 × s.e.m. calculated from the measurement of 100 to 300 ratios. ⁸⁷Sr/⁸⁶Sr ratios were normalized to ⁸⁶Sr/⁸⁸Sr = 0.1194. Twelve analyses of National Bureau of Standards-987 were completed during the course of this experiment (mean ⁸⁷Sr/⁸⁶Sr = 0.710215 ± 0.000062 (δ⁸⁷Sr = 1.4594 ± 0.088‰)). Rearing sites that share a common superscript letter are not significantly different (Fisher's LSD multiple comparison, P < 0.05). Water and fish values were pooled for analyses, excluding two fish (*) that incorporated a significant portion of their signal outside the stream in which they were collected.

1. Koch, P. L. *et al. Earth Planet Sci. Lett.* **108**, 277–287 (1992).
2. Fisher, R. S. & Stueber, A. M. *Wat. Resour. Res.* **12**, 1061–1068 (1976).
3. Graustein, W. C. & Armstrong, R. L. *Science* **219**, 289–292 (1983).
4. Åberg, G., Jacks, G. & Hamilton, P. J. *J. Hydrol.* **109**, 65–78 (1989).
5. Miller, E. K., Blum, J. D. & Friedland, A. J. *Nature* **362**, 438–441 (1993).
6. Bailey, S. W., Hornbeck, J. W., Driscoll, C. T. & Gaudette, H. E. *Wat. Resour. Res.* **32**, 707–719 (1996).
7. Simkiss, K. in *The Ageing of Fish* (ed. Bagenal, T. B.) 1–12 (Unwin, Old Woking, 1974).
8. Graustein, W. C. in *Stable Isotopes in Ecological Research* (eds Rundel, P. W., Ehleringer, J. R. & Nagy, K. A.) 491–512 (Springer, New York, 1989).
9. Kalish, J. M. *Fish. Bull. US* **88**, 657–666 (1989).
10. Limburg, K. E. *Mar. Ecol. Prog. Ser.* **119**, 25–35 (1995).
11. Chamberlain, C. P. *et al. Oecologia* **109**, 132–141 (1997).
12. Campana, S. E. & Neilson, J. D. *Can. J. Fish. Aquat. Sci.* **42**, 1014–1032 (1985).
13. Hodell, D. A., Mueller, P. A., McKenzie, J. A. & Mead, G. A. *Earth Planet. Sci. Lett.* **92**, 165–178 (1989).

DNA fingerprints from fingerprints

Forensic scientists regularly generate genetic profiles from old blood stains, seminal stains, vaginal swabs, hair, bone, urine and cigarette butts^{1–6}. We show that an individual's genetic profile can now also be generated from swabs taken from objects touched by hands, providing a new tool for crime scene investigations. Our findings also demonstrate the need for caution when handling exhibits and when interpreting results.

We swabbed specific areas of hands and objects with cotton cloth dampened with sterile water, using disposable forceps. We extracted⁷ and quantified (ACES 2.0+ program, Gibco BRL) DNA from these swabs, and typed for a short tandem repeat locus using the polymerase chain reaction⁷. We compared the results with independent typings of blood or buccal samples from participating individuals.

Initial tests showed that we could readily obtain correct genetic profiles from swabs taken directly from the palm of a hand (13 of 13). DNA yields varied from 2 to 150 ng (average 48.6 ng). Dry hands and those that had been washed recently tended to provide the least DNA.

Swabs of objects handled regularly by specific individuals all provided genetic typings that matched the user. Objects included: leather briefcase handles ($n=3$, mean 75 ng DNA), pens ($n=3$, mean 1.6 ng), a car key ($n=1$, 1.1 ng), a personal locker handle ($n=1$, 3.7 ng) and telephone handsets ($n=5$, mean 10.3 ng). One of the telephone handsets also clearly displayed the genetic profile of a known secondary (minor) user.

Furthermore, a number of pre-cleaned objects held for a relatively short period of time (15 min) including: plastic knife han-

dles ($n=6$, mean 17.8 ng DNA), a mug ($n=1$, 6.8 ng), a glass ($n=1$, 34 ng) as well as new vinyl gloves worn for 20 to 90 min ($n=8$, mean 51 ng) gave the genetic profile of the holder or wearer. We found alleles in addition to those of the wearer in samples from two of the gloves, which could be due to secondary transfer. We also found that swabs of the inside of worn (1 min) condoms ($n=4$), where no ejaculation occurred, also provide the wearer's genetic profile (mean 11.2 ng DNA). This is relevant to some sexual offence investigations.

DNA yields from swabs of polypropylene tubes held for varying lengths of time (5 s, 30 s, 3 min, 10 min), did not differ significantly, indicating that substantial transfer of material occurs during initial contact.

Objects handled by many individuals all produced profiles with multiple alleles of varying intensity. To determine the effect of multiple handlers, we exchanged polypropylene tubes between individuals (2 or 3, 10 min each) with different genotypes. Although the material left by the last holder was usually present on the tube, that of previous holders was also retrieved to varying extents. The strongest profile obtained was not always that of the person who last held the object, but was dependent on the individual. We regularly observed profiles of previous holders of a tube from swabs of hands involved in these exchanges, showing that in some cases material from which DNA can be retrieved is transferred from object to hand (secondary transfer).

Also, hands swabbed before and after a one-minute handshake revealed the transfer of DNA from one individual to another in one of the four hands tested. Thus genetic profiles from objects handled by several people or from minute blood stains on touched objects may be difficult to interpret.

There are many cases in which the genetic profile of individuals who may have handled or touched particular objects associated with a crime could be extremely important to an investigation. Our methods have already been used at our laboratory to provide evidence in attempted murder, rape, armed robbery, extortion and drug-trafficking cases.

Roland A. H. van Oorschot
Maxwell K. Jones

Victoria Forensic Science Centre,
Victoria Police, Macleod,
Victoria 3085, Australia

1. van Oorschot, R. A. H., Gutowski, S. J., Robinson, S. L., Hedley, J. A. & Andrew, I. R. *J. Forensic Sci.* **41**, 142–145 (1996).
2. Higuchi, R., von Beroldingen, C. H., Sensabaugh, G. F. & Erlich, H. A. *Nature* **332**, 543–546 (1988).
3. Hagelberg, E., Sykes, B. & Hedges, R. *Nature* **342**, 485 (1989).
4. Hagelberg, E., Gray, I. C. & Jeffreys, A. J. *Nature* **352**, 427–429 (1991).
5. Brinkman, B., Rand, S. & Bajajowski, T. *Int. J. Legal Med.* **105**, 59–61 (1992).
6. Hochmeister, M. N. *et al. Int. J. Legal Med.* **104**, 229–233 (1991).
7. van Oorschot, R. A. H., Gutowski, S. J. & Robinson, S. L. *Int. J. Legal Med.* **107**, 121–126 (1994).

How many replicons make a nodule?

Allan Downie, in his commendable News and Views discussion¹ of our work², suggests that a second, non-chromosomal symbiotic replicon could be present in the bacterial *Rhizobium* species NGR234. So far, our as-yet unpublished work on physical mapping and random sequencing of the NGR234 genome by V. Viprey, C. Freiberg and X. P. has produced no evidence of another plasmid. Further, the electrophoretic methods we used could visualize the twin symbiotic plasmids (relative molecular mass (M_r) 1×10^9) of *R. meliloti*, but in the event consistently demonstrated only a single plasmid of M_r $3.1 \times 10^8 \pm 2 \times 10^7$ in NGR234 (refs 3, 4). Thus, although Morrison *et al.*⁵ reported that NGR234 contains two plasmids of M_r 3×10^8 and 4.5×10^8 , it seems unlikely that a second plasmid exists. It is true, however, that some nodulation and other symbiotic genes map to the chromosome⁶. (Incidentally, the *fixABC* cluster is present on pNGR234a.)

Xavier Perret, William Broughton

LBMPs, Université de Genève,
1 chemin de l'Impératrice,
1292 Chambésy/Genève, Switzerland
e-mail: broughtw@sc2a.unige.ch

1. Downie, A. *Nature* **387**, 352–354 (1997).
2. Freiberg, C. *et al. Nature* **387**, 394–401 (1997).
3. Broughton, W. J., Heyck, N., Meyer, Z. A., H. & Pankhurst, C. E. *Proc. Natl Acad. Sci. USA* **81**, 3093–3097 (1984).
4. Broughton, W. J. *et al. Arch. Microbiol.* **141**, 14–21 (1985).
5. Morrison, N. A. *et al. J. Bacteriol.* **160**, 483–487 (1984).
6. Perret, X., Broughton, W. J. & Brenner, S. *Proc. Natl Acad. Sci. USA* **88**, 1923–1927 (1991).

Unique morphology of the human eye

Human eyes have a widely exposed white sclera surrounding the darker coloured iris, making it easy to discern the direction in which they are looking¹. We compared the external morphology of primate eyes in nearly half of all primate species, and show that this feature is uniquely human. Humans have the largest ratio of exposed sclera in the eye outline, which itself is elongated horizontally. We suggest that these are adaptations to extend the visual field by allowing greater eye movement, especially in the horizontal direction, and to enhance the ease of detecting the gaze direction of another individual.

We measured three parameters in 88 primate species: an index of exposed sclera size (SSI) in the eye outline, the width–height ratio (WHR) of the eye outline and the sclera coloration. Human eyes have the largest

SSI and the outline shows extraordinary horizontal elongation. Humans are also the only species with white sclera. We failed to detect any significant sexual or racial differences in these parameters. Although a small number of primates had pale sclera (*Macaca sylvanus*, *M. nemestrina*) or brown sclera with small white regions to the side of the iris (*Saguinus midas*, *S. labiatus*, *Callithrix argentata*, *Callimico goeldii*), almost all other primates examined have similar coloration to that of the skin around the eyes (Fig. 1).

SSI correlates with weight and crown–rump length^{2,3} ($r=0.59$, $P<0.001$ in both cases), sitting height ($r=0.65$, $P<0.001$) and walking height ($r=0.71$, $P<0.001$) (Fig. 2). Larger SSIs allow the iris a wider range of movement and hence a larger visual field. This may be advantageous to larger species where eye movement becomes increasingly more efficient than head or body movement, especially as comparative eyeball size is smaller in larger animals⁴. In addition, in small species with comparatively large eyeballs in a small skull, muscle space may be seriously limited.

We video-recorded various primates eating (18 species, 29 individuals), and counted eye and head movements. The amount of scanning by eye movement alone was correlated with SSI ($r=0.73$, $P<0.005$) and was high in humans ($61 \pm 28\%$ of horizontal scanning, $n=5$) compared with other primates (4.3–24.4%; mean, 10.6%). The highest rate in non-human primates was observed in the chimpanzee, *Pan troglodytes* (20–35%, $n=3$).

WHR is largest in terrestrial species, smallest in arboreal species, with semi-arboreal species lying in-between (Fig. 1). There is a corresponding high ratio of horizontal to vertical scanning in terrestrial species, as might be expected to suit this lifestyle, and a low ratio in arboreal species. The ratio is correlated with WHR (scanning time ratio: $r=0.74$, $P<0.001$; frequency ratio: $r=0.88$, $P<0.001$).

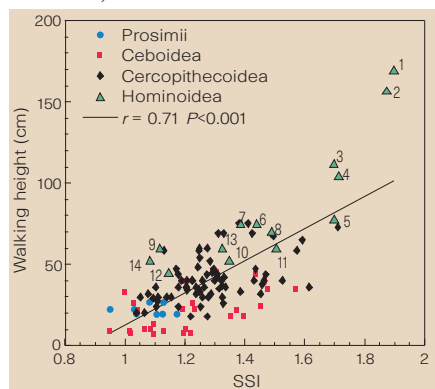


Figure 2 Relationship between SSI and walking height. 1, Human (Japanese) male; 2, female; 3, *G. gorilla* male; 4, female; 5, *P. pygmaeus* female; 6, *Pan troglodytes* male; 7, female; 8, *P. paniscus* male; 9, *H. agilis* female; 10, *H. lar* male; 11, female; 12, *H. pileatus* female; 13, *H. syndactylus* male; 14, female.

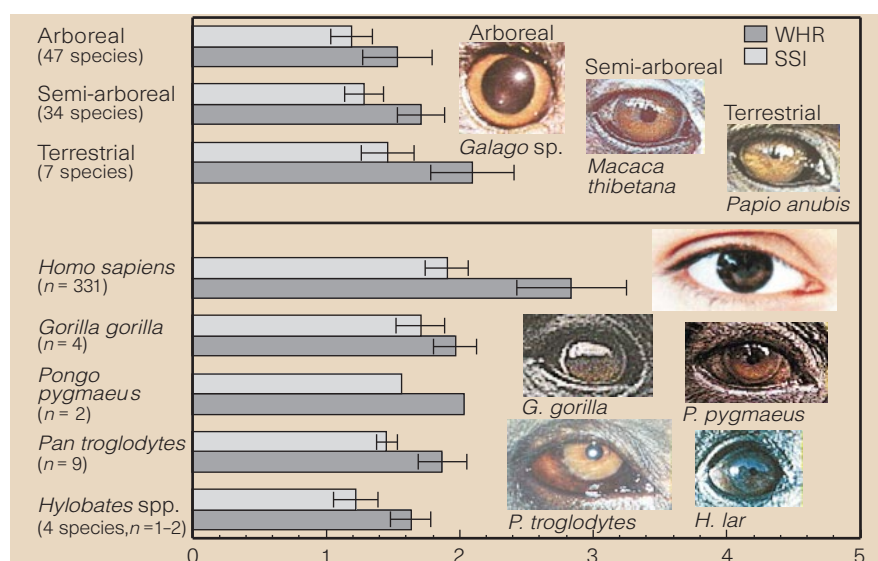


Figure 1 Variation of WHR and SSI (mean $\pm$ s.d.). Difference between habitat types was significant (WHR: $F_{2,85} = 18.69$, $P<0.01$, least significant difference, mean square of errors = 0.058, $P<0.01$; SSI: $F_{2,85} = 10.86$, $P<0.01$, least significant difference, mean square of errors = 0.024, $P<0.01$). We studied frontal full-face images without obvious facial expression of 387 adult animals (88 species: Prosimii, 10; Ceboidea, 26; Cercopithecoidea, 43; Hominoidea, 9). Facial images of 80 species were recorded by video camera at the Japan Monkey Centre and those of 8 species (*Microcebus* sp., *Loris tardigradus*, *Perodicticus potto*, *Tarsius* sp., *Saguinus imperator*, *Pithecia monachus*, *Cacajao rubicundus*, *Cercopithecus hamlyni*) were collected from books. 182 Japanese, 80 Caucasian⁸ and 68 Afro-Caribbean^{8,9} adults were observed. Images were analysed using the NIH Image program. WHR = distance between the corners of the eye/longest perpendicular line between the upper and lower eyelid; SSI = width of exposed eyeball/diameter of iris.

Microscopic analysis of Japanese macaque (*Macaca fuscata*) eyes showed brown pigmentation of the sclera tissue around the cornea, apparently common in primates and other mammals. This pigmentation was thought to reduce glare as it is absent in many nocturnal and crepuscular species⁵, but nocturnal primates (*Gelago senegalensis*, *Tarsius syrichta*, *Perodicticus potto*, *Nycticebus coucang* and *Aotus trivirgatus*) also had coloured sclera and diurnal humans showed no pigmentation.

In many primates, gaze direction is important in communication, and direct eye contact often elicits attacks. Sclera pigmentation to obscure the gaze direction may thus be adaptive⁶. It may also serve to deceive natural predators, as if the predator believes that the prey animal is aware of its presence, it may be less likely to attack⁷. In some nonhuman primates (9 of 10 species examined), sclera coloration of newborns was paler than adults, indicating that infant gaze signals might have special meanings in these species. In all of 14 species examined, including humans, SSI and WHR of newborns were lower than those of adults.

The human sclera is much paler than the facial skin or iris so it is very easy to discern the gaze direction. Predation risk might have decreased with the evolution of enlarged body size and the use of tools and fire. In addition, gaze-signal enhancement might aid the communication required for increased cooperative and mutualistic behaviours to allow group hunting and

scavenging. A small change in sclera coloration may have altered 'gaze-camouflaged' to 'gaze-signalling' eyes. SSI and WHR of human eyes are even larger than those of gorillas, the largest primate, which suggests adaptation for gaze-signal enhancement. The uniqueness of human eye morphology among primates illustrates the difference between humans and other primates in the ability to communicate using gaze signals.

Hiromi Kobayashi, Shiro Kohshima

Basic Biology, Faculty of Bioscience and Biotechnology (c/o Faculty of Science), Tokyo Institute of Technology, 12-1, O-okayama 2-chome, Meguro-ku, Tokyo 152, Japan
e-mail: hiromi@chem.titech.ac.jp

1. Morris, D. *Body Watching* (Equinox, Oxford, 1985).
2. Itani, J. & Uehara, S. *The Encyclopaedia of Animals* (ed. Macdonald, D. W.) (Heibonsha, Tokyo, 1986).
3. Napier, J. R. & Napier, P. H. *The Natural History of the Primates* (MIT Press, Cambridge, MA, 1985).
4. Schultz, A. H. *Am. J. Phys. Anthropol. Old Ser.* **26**, 389–408 (1940).
5. Duke-Elder, S. S. in *System of Ophthalmology* (ed. Duke-Elder, S. S.) 453 (Kimpton, London, 1958).
6. Perrett, D. I. & Mistlin, A. J. in *Comparative Perception — Complex Signals* (eds Stebbins, W. C. & Berkley, M. A.) 187–215 (Wiley, New York, 1990).
7. Sherman, P. W. *Science* **197**, 1246–1253 (1977).
8. Ohara, K. *ONE* (Tsukui Shokan, Tokyo, 1970).
9. Gomi, A. *Americans 1.0 1994 Los Angeles* (Fuga Shobo, Tokyo, 1994).

erratum

In the Scientific Correspondence "Evidence for stone age cranial surgery" by Kurt W. Alt *et al.* (*Nature* **387**, 360; 1997), the carbon-14 estimate of the age of the human bones was printed incorrectly. It should have read 'Utrecht ¹⁴C laboratory sample UIC-5406: 6,155 $\pm$ 39 radiocarbon years before present, ~5100 BC'.

Politics of eighteenth-century science

The Correspondence of John Flamsteed, First Astronomer Royal: Volume II, 1682–1703

edited by Eric G. Forbes, Lesley Murdin and Frances Wilmoth

Institute of Physics Publishing: 1997. Pp. 1,095. £140, \$280

Owen Gingerich

"Tycho Brahe was borne an hundred yeares before me," the first Astronomer Royal wrote in 1700, "His observatory built 100 yeares before ours. He labored about 25 yeares on his Catalogue, had a large estate of his owne with about 3000 dollars per Annum allowance from the King of Denmar[k] which in that cheap country and in those times was more than 3000 sterling here. He had 8 or 10 assistants. Yet with his large apparatus of Instruments and helpe he left us not more than 780 places of fixed stars.... When I sat down here I had onely £100 per Annum allowance from the Office of Ordnance, with a surly, silly laborer.... I was promised a Murall Instrument, but a freind of ours [Robert Hooke] contrived one so that it wa[s] useless to me.... Finding a large and weighty Instrument necessary, I contrived and began a new Murall Arch, but my servant died before the ribbs were joyned together.... I finisht it.... Now I have by me about 1800 places of fixed stars rectified in my Catalogue."

In capsule form, John Flamsteed has here given much of the principal story found in this second instalment of his correspondence (for a review of the first volume, see *Nature* 380, 212; 1996). The letters in this volume begin in 1682, when the 35-year-old Flamsteed had been at Greenwich for seven years. For a decade the most voluminous exchange is with an Irish mathematician, William Molyneux; they discuss, among other things, optics, tides and Newton's *Principia*. Then, very suddenly, Molyneux incurs the Astronomer Royal's displeasure, apparently by not giving Flamsteed any credit in his book on optics, and perhaps also because Flamsteed's rival and irritant Edmond Halley had assisted in its forthcoming publication. Like a guillotine, Flamsteed's wrath descends and the (for us, rather tedious) correspondence abruptly ends.

Towards the end of this vast tome, a lively exchange ensues between Flamsteed and Abraham Sharp, a gifted geometer and instrument maker and his former observing assistant, the man who eventually assisted in bringing the magnificent Flamsteed atlas into reality after the Astronomer Royal's death in 1719. But that part of the story belongs to volume 3, which will pick up in 1703 where this volume leaves off.

In between Molyneux and Sharp lies the vigorous and fascinating exchange with Isaac

Newton, for the most part already available in volumes 2–4 of *The Correspondence of Isaac Newton*, but here more fully placed in the context of the Astronomer Royal's life and work. What eventually becomes (in Robert Oppenheimer's memorable phrase in another context) the dance of two scorpions in a bottle, starts out civilly enough. Flamsteed had seen the sheets of the *Principia* well before publication. Newton had begun the correspondence a few years earlier, asking first for positions of a few stars, and later for positions of the comet of 1680, which he now admitted to be but one comet as Flamsteed had previously insisted, without conceding any credit to the Astronomer Royal. Not until four years after the appearance of the *Principia* did the exchange resume, with Newton pressing for the publication of Flamsteed's star catalogue, and then, with increasing insistence, for lunar positions. These Flamsteed supplied with increasing reluctance and with demands that they not be published without his strict permission, and finally Newton in some frustration declares: "And for my part, I am of the opinion that for your Observations to come abroad thus with a *Theory which you ushered into the world and which by their means has been made exact would be much more for their advantage and your reputation than to keep them private till you dye or publish them without such Theory to recommend them*. For such a Theory will be a demonstration of their exactness and make you readily

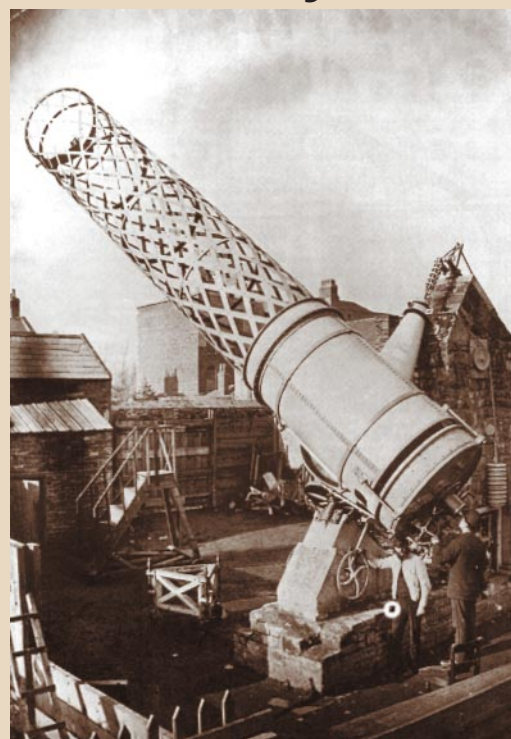
acknowledged the Exactest Observer that has hitherto appeared in the world."

No letters bring news of Halley's complaints that the Astronomer Royal has been withholding observations that Newton required for perfecting his lunar theory — Flamsteed must have heard them from the gossip in the London coffeehouses — but his letters bristle with allusions to them, and the missive that opens this review is part of his not unreasonable defence. In 1700 Flamsteed wrote that Newton is "a good Man at the Bottom, but through his Naturall temper suspicious", but he seldom had anything good to say about Halley. Flamsteed was particularly pleased when he saw a possible opportunity to unmask what he perceived as Halley's perfidy: "I desire not to make him smart but blush," he wrote late in 1700.

This fascinating window into eighteenth-century scientific politics leaves us with some sympathy for the unbending, strait-laced Flamsteed. His cavils against David Gregory's refusal to believe that he had measured the size of stars — 15" for Sirius — or his stellar parallax make us more ambivalent; we have to look elsewhere to find that Flamsteed had apparently stumbled onto aberration, but neither he nor Gregory, Halley or Newton got the interpretation right.

We need to stay tuned for the climax of these conflicts, and the sordid story of how the star catalogue got printed, in the third and final volume of this valuable series. The

Stars in their eyes



For 95 years Thomas and Howard Grubb, father and son, supplied astronomical instruments to the world. Ian Glass of the South African Astronomical Observatory has used many of the Grubbs' original letters and documents to write the history of their Dublin-based company from its beginnings in 1830 as a billiard table manufacturer to its position as maker of some of the largest and best known telescopes of the Victorian era. The company was responsible for "some of the world's finest telescopes, including some which are still in full operation", writes the astronomer Patrick Moore in the foreword to the book. *Victorian Telescope Makers: The Lives and Letters of Thomas and Howard Grubb* is published by Institute of Physics Publishing (£30, \$50). Left: the Great Melbourne Telescope assembled in Grubb's yard in Dublin before its shipment to Australia in 1868. The instrument was described at the time as "a great triumph of mechanical engineering and optical skill".

editors (but not this reviewer) have fastidiously followed Flamsteed's idiosyncratic punctuation, have added translations for the occasional Latin letters, and have provided thousands of annotations. □

Owen Gingerich is an historian of astronomy at the Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA.

Cleaning devils

Biting the Dust: The Joys of Housework

by Margaret Horsfield
Fourth Estate: 1997. Pp. 292. £14.99

Richard Davenport-Hines

"Our idea of dirt is compounded of two things, care for hygiene and respect for conventions", Mary Douglas defined in her superb book on pollution and taboo, *Purity and Danger* (1966). Dirt represents disorder in Douglas's analysis, although Christian tradition also identifies it with vice: "Wash me clean of my guilt, purify me from my sin", the psalmist said. The sanitation campaigns of nineteenth-century reformers such as Charles Kingsley met a desperate human need, but it was not long before consumerism exploited the human impulses against dirt, disorder and guilt. As Henry Ward Beecher commented in 1885, "If cleanliness is next to Godliness, soap must be considered as a means of grace."

With the exception of a relatively few public health campaigners and civil engineers, dirt has not been a masculine preoccupation. The fight against dirt has been women's work, and, as Margaret Horsfield shows in her lively, discursive reflections on housework, if the great victories of this war have been the subject of justified pride, its petty skirmishes have often constituted a cruel oppression or mad obsession. Despite the mechanization of housework in the industrialized world since 1950, domestic chores have for some women been not merely a matter of necessary daily routine but have become a dominant psychological struggle against evil.

Horsfield is a radio journalist rather than an anthropologist or historian. Although she gives heartbreaking accounts of the workload of Victorian housemaids, and fascinating examples of early advertising jingles such as "The Happy Housewife's Song" of 1899, she has no ambition to supplant Caroline Davidson's authoritative history of British housework from 1650 to 1950, *A Woman's Work is Never Done* (1982). Horsfield's method in compiling this account of housework since the nineteenth century has been that of the sympathetic radio journalist; it is filled with short, apt quotations from women sharing "their dark secrets of dirty ovens, glowing accounts of gleaming floors and true confessions of reeking dishtcloths". She has garnered



"A happy washerwoman", c. 1901

references to housework from a delightfully wide range of novelists. Some little-known byways of hygiene history are explored: Horsfield gives an interesting sketch of an organization called the Cleanliness Institute set up in 1927 by US soap manufacturers worried by falling sales which indoctrinated housewives with excessive new standards of domestic cleaning. She quotes a conclusion from Vincent Vinikas's specialist study *Soft Soap, Hard Sell: American Hygiene in an Age of Advertisement* (1992): "Manufacturers had to let Americans know, not just that they were still soiled, but that they could never be sanitary enough. As the country became cleaner than ever before, manufacturers had to dig up dirt."

Horsfield is more anecdotal than analytical, and some of the anecdotes are perhaps less charming than she realizes. There are several super-hygienic little children (apparently all from North America, although this is not explicit) whose eccentricities are not particularly winning. Horsfield describes herself as "a Canadian brought up in a cold climate in a clean house in a puritanical society" by a mother "who suffers, often loudly, if a house is dirty" and makes her daughter "feel like a lazy slut". This cultural background dominates the book. One chapter is devoted to the ways in which some North American mothers and daughters torture one another with "utterly mad" controls, reproaches, rebellion and guilt about housework. It is not clear whether the appalling emotional blackmail recounted in this chapter is peculiarly American, but the overall impression is painfully mean-spirited. Although every reader will sympathize with Horsfield's strong reaction and compulsive scrubbing after her home was vandalized by burglars, she is too scornful of the "young male renovator of Thatcherite persuasions" who chose dark blue bathroom fittings for a

London flat heedless of the white stains that would be left by chalky London water. Horsfield has little tolerance for people who put aesthetics first.

The diverse European approaches to housework are a relief after the monolithic puritan earnestness of the North Americans. Horsfield rather begrudgingly quotes the French feminist Simone de Beauvoir asserting in *The Second Sex* (1952) that "the rage for cleanliness is highest in Holland, where the women are cold, and in puritanical civilisations, which oppose an ideal of neatness and purity to the joys of the flesh." De Beauvoir preferred southern voluptuousness with its concomitant dirt: "If the Mediterranean Midi lives in a state of joyous filth, it is not only because water is scarce there: love of the flesh and its animality is conducive to toleration of human odour, dirt, and even vermin." The admirable Germans have coined two words to describe precisely the potential excesses of housework: *Putzimmel*, meaning a chronic frenzy of cleaning, and *Putzteufel*, meaning someone who is a cleaning devil. Horsfield certainly fills her readers with pity and horror at the cleaning devils. □

Richard Davenport-Hines is at 51 Elsham Road, London W14 8HD, UK.

Knowing our minds

The Culture of Education

by Jerome Bruner
Harvard University Press: 1996. Pp. 224.
\$24.95, £9.95

Liam Hudson

Jerome Bruner is that rarity among psychologists: a man who has had a distinguished career in conventional terms — a professor in his time at both Harvard and Oxford — yet who has gazed out intently towards the unfashionable world of school education. In the 1960s, he propounded the view that "any subject could be taught to any child at any age in some form that is honest", so bringing together theories of child development with the analysis of an academic discipline's essentials, its intellectual core as opposed to its trappings. He did this by dexterous use of the Heraclitean spiral. The appeal of the "spiral curriculum" to educational reformers was profound, as it encompassed aspects of the mantra and the heuristic. One regrets only that it proved so difficult to apply.

Since those heady days, something has plainly gone wrong with both schooling and psychology. Teaching methods conceived as progressive have seemed to deny whole generations of underprivileged children the survival skills of reading, writing and calculation. Even in leafier suburbs of the United States, children rendered jittery by regimes of compulsory togetherness are being diagnosed as suffering from a 'disorder', and are

being dosed with Ritalin.

Meanwhile, psychology has grown and grown. (The back-of-an-envelope calculations are frightening. There are at least 100,000 professional psychologists in the world at the moment. If each of us earns US\$25,000 a year, our discipline is costing the world's economies something in the order of \$25 billion a decade in salaries alone.) But what unchallengeable discoveries about human nature have resulted? Most of us struggle to name one.

Despite changed times, the tone of the nine recent essays in *The Culture of Education* remains optimistic, humane. While acknowledging that "psychology, alas, seems to have lost its center and its great integrating questions", Bruner believes that it has "given up prematurely". He proposes an "exemplary subject-matter area" designed to give back to the discipline its coherence and élan. The example he chooses is that of "other minds", of "intersubjectivity"; and he suggests that this perennially alluring problem should be approached in the context of the study of infant-mother interaction.

The choice is judicious in that it re-establishes psychology as the science of minds. But it also invites criticism. His proposal could be seen as taking one of academic psychology's more recent bandwagons and assuming that what it trundles along is a Royal Road; and, in doing so, citing a handful of performances, some of them in all conscience humdrum, and awarding them canonical status. Those psychoanalysts and psychotherapists who, for generations, have explored the nature and complexities of empathetic communication may also feel unjustly ignored.

There is, too, the question of pertinence. The bearing on the psychology of adult life of findings about eye-contact between mother and infant remains to be demonstrated, for example. A demonstration can be achieved only by long-term follow-up studies, or by the retrospective study of adults. It is hard, after all, to envisage Newton's gifts as dependent on an empathetic bond with his mother or teacher. On the contrary, it is tempting to see imaginative activity as fuelled by just those deep-seated dissonances that an empathetic educational regime might seek to resolve.

As his title indicates, Bruner now views the school as a vehicle of 'culture'. What he says is civilized and kindly, but he does little to quiet the anxieties of those who see the values inadvertently transmitted in schools as the values of consumerism: the will to explore life's headier excitements in terms that happen to enhance others' cash flows. Or of those, by no means all Marxists, who see school as acting, despite the best intentions, as the recruiting ground for an under-class.

As in everything that Bruner writes, though, there is the stamp of imaginative energy, an exceptional breadth of reading and

subtle thought. The most arresting of the essays, "The narrative construal of reality", chronicles work on story-telling undertaken with his wife Carol Fleisher Feldman. This centres on a list. "It is through our own narratives that we principally construct a version of ourselves in the world, and it is through its narrative that a culture provides models of identity and agency to its members", he writes. Bruner identifies the nine "universals" of such story-telling. These range from "a structure of committed time", through "the centrality of trouble" and "inherent negotiability" to narrative's "historical extensibility". Each is discussed in turn.

While the persuasive force of the 'spiral curriculum' lay in its elegance, it may be in the apparent awkwardness of such taxonomies that the prospect of a more enduring advance lies. This is especially true when, as in this case, the analysis bears directly on points where psychologists' preoccupations overlap with those of their neighbours: sociologists and political scientists, literary critics, historians, philosophers of science, biographers, psychotherapists, management consultants, paediatricians and, as Bruner himself insists, brain scientists.

This, then, is a text within which there is evidence of fruitful tension; of the author's left hand operating independently of his right. Implicitly, and despite his professed intentions to reaffirm the educator's mission and to rebuild psychology's centre, the most penetrating passages of his argument help to make the case for decentralization: for disciplines alive with excitement at their boundaries but with their heartlands unexplored. Even for sets of disciplines with no habitable centres whatever. □

Liam Hudson is at Balas Copartnership, 34 North Park, Gerrards Cross, Buckinghamshire SL9 8JN, UK.

The latest C-theory

Complexity: Hierarchical Structures and Scaling in Physics

by Remo Badii and Antonio Politi
Cambridge University Press: 1997. Pp. 318.
\$74.95, £50

Steven H. Strogatz

Complex systems consist of an enormous number of simple units whose interactions can lead to unexpected collective behaviour. Think of an ant colony that seems to have a mind of its own, or the blackouts that plague the electrical power grid of the western United States each summer. The dream of complexity theory is to discover the laws governing such complex systems, whether they be ecosystems, the weather or corporations in the marketplace.

Many people first got wind of complexity theory in Michael Crichton's novel *The Lost*

World (Random House, 1995). As Crichton explains:

The scientists came from many fields — physics, economics, biology, computer science. What they had in common was a belief that the complexity of the world concealed an underlying order which had previously eluded science, and which would be revealed by chaos theory, now known as complexity theory. In the words of one, complexity theory was "the science of the twenty-first century."

Some sceptics will grumble that they that we have heard all this before. First it was cybernetics in the 1960s, then catastrophe theory in the 1970s, then chaos theory in the 1980s. Now complexity theory is the latest grandiose C-theory that purports to be the science of the next century.

To gain a more balanced view about the prospects for a successful theory of complexity, one first needs to appreciate what is known, and unknown, about the behaviour of nonlinear systems. (Linear systems cannot be complex; the principle of superposition ensures that the whole behaves in precisely the same way as the sum of the parts. This is why linear dynamical systems are completely solvable, and also why they are incapable of any surprises.) The 1980s were the boom years for the study of nonlinear dynamics, and chaos in particular. Mathematicians and physicists studied the simplest nonlinear models: ordinary differential equations and discrete-time iterated maps with only a handful of state variables. The seminal discovery, due to Mitchell Feigenbaum and others, was that there are certain universal laws governing the transition from regular to chaotic behaviour; roughly speaking, completely different physical systems can become chaotic in the same way. The theory was tested and confirmed in experiments on systems as diverse as convecting fluids, pulsating chick heart cells, semiconductors, electronic circuits and chemical oscillators.

But the limitation of chaos theory is that it applies only to systems with a few active degrees of freedom, or, equivalently, systems so heavily constrained that they act effectively as if their state spaces were low-dimensional. As soon as many degrees of freedom become active in both space and time, as in a turbulent fluid or a fibrillating heart, the results of elementary chaos theory do not apply. Enter complexity theory — except that complexity theory does not exist yet. In fact, no one even knows how to define complexity, let alone deduce its underlying laws.

The main goal of this graduate monograph is to characterize complexity. Besides ideas from nonlinear dynamics, the authors consider relevant measures of complexity proposed in computer science (for example computational complexity, the Chomsky hierarchy in formal language theory), cellu-

lar automata (Stephen Wolfram's classification and modifications thereof), physics (statistical mechanics, spin glasses, hierarchical structures) and probability (different notions of mixing in ergodic theory). But rather than presenting a hodgepodge of ideas, the authors have neatly organized the discussion around a common theme: the symbolic dynamics of a one-dimensional, stationary symbol sequence. They argue that many interesting problems fall into this framework, and use it as a test case to illustrate the various complexity measures.

The approach throughout is no-nonsense. You won't find any bedtime reading about consciousness or economics here. There is no mention of more ambitious but speculative views of complexity, such as Per Bak's self-organized criticality or Stuart Kauffman's evolution on rugged fitness landscapes. But for a mathematically mature reader determined to ponder what complexity might mean, this book will provide some rewarding exercise. □

Steven H. Strogatz is in the Department of Theoretical and Applied Mechanics and the Center for Applied Mathematics, Cornell University, Ithaca, New York 14853, USA.

Truth buried in history

This is Biology: The Science of the Living World

by Ernst Mayr

Harvard University Press: 1997. Pp. 327.

\$29.95, £19.95.

David Baltimore

In this wide-ranging book, Ernst Mayr, one of the doyens of evolutionary biology, raises many important questions about the nature of biological research. He examines them in a scholarly yet approachable way. It is not an easy book to read, but it is rewarding.

A key question for Mayr is why physics is so different from biology. He emphasizes the purely historical aspect of biology absent in physics: physical events follow from a few laws and facts whereas biological systems are the product of natural selection that has chance as its substrate. To explain why any particular biological system exists at any one time requires constructing a 'historical narrative' that cannot be proved to be true because its truth is buried in history.

This argument, while basically sound, leaves out an important aspect of biology: even though any particular bit of biology is the product of evolution, it is also lawful. After all, one can do experiments and get reproducible results. Although Mayr does distinguish between these two aspects of biology — he recognizes that biological systems must obey the laws of physics and chemistry — this law-

ful component is to him less interesting than the evolutionary component. As a molecular biologist, I am tempted to say that he has missed the point of the past 50 years in biology. I am persuaded, however, that this is not the case. Rather, he is more comfortable fighting the battles of evolutionary biology in their pure state.

Mayr takes the philosophers of science to task for forming their analyses of science from examples in physics and then generalizing to all of science. He argues that the historical side of biology requires it to have a different philosophy. Having grown up reading the logical positivists and wondering why they made no sense, I too have a sympathy for this view. Mayr puts Thomas Kuhn in the camp of the physically oriented philosophers. Although Kuhn does focus on physics, he was to me the first philosopher of science who made sense. Kuhn's notion of paradigm shifts particularly bothers Mayr. But, having lived through some authentic paradigm shifts in biology (maybe even helping to create one), I believe that here Mayr becomes myopic in concentrating only on the evolutionary side of biology and ignoring the lawful side of molecular biology where paradigms regularly rise and fall.

Mayr often emphasizes 'genetic programs' as the basis of the functioning of biological systems but he never carries his treatment of them very far. I think he is right that a major contribution of modern biology to evolutionary thinking is to see the changes over evolutionary time more as changes in programs of genetic expression than as changes in structural genes. It would be interesting to know from Mayr how this changed perspective influences evolutionary theory.

There is much more in the book, so much that at times it loses focus and becomes a stream of discussion. Not a gently flowing stream either: it plunges into many arguments in evolutionary biology, outlining the development of concepts and Mayr's present views — but with few examples from the natural world.

Mayr comes across as a man of fine, liberal sentiments who has lived through the wars of modern evolutionary theory and now enjoys reflecting on them. He ends with a chapter called "Can evolution account for ethics?" in which he comes down on the side of "evolutionary humanism", saying that every person "shares a responsibility for the future of our species" which is "as much a part of cultural ethics as concern for the individual". This is Mayr's shot at the ethics of the selfish gene because he believes that arguments for altruism based on shared genes are insufficient to generate ethical norms in an interdependent society. He never deals with the nasty arguments against group selection which have always left me frustrated because my instinct agrees with him — there must be group selection, but I have not yet seen the way around selectivist arguments in favour of the gene as the unit of selection.

This is a book designed to make one think. Difficult to penetrate in places, and sometimes fairly technical, its scope encompasses deep issues. Particularly enjoyable is the chapter on human evolution, which seems to me (a rank amateur in this part of biology) thoughtful and illuminating. Mayr raises the fascinating question of how we humans have been able to change our society so remarkably in the past thousands of years — occupying many niches of climate and geography — without much change in our gene pool. It is just one of many unanswered questions that course through his fertile brain and have found an outlet in this volume. □

David Baltimore is at the Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA.

Corrections

● In his review of *Who Gave Pinta to the Santa Maria? Tropical Diseases in a Temperate Climate* by Robert S. Desowitz (Nature 387, 36; 1997), Len Goodwin says that the book has no index. This is true only of the advance proof copy he was sent for review, not the finished version of the book.

● In Jeremy Hyams's review of *Life Itself* (Nature 386, 778; 1997), the book's author, Boyce Rensberger, is incorrectly described as having attended the physiology course at Woods Hole Oceanographic Institute in Massachusetts, rather than the Marine Biological Laboratory. The laboratory has held a physiology summer course in Woods Hole for over a hundred years, and has been operating a science-writing fellowship programme since 1986, of which Rensberger now serves as co-director.

1997 Rhône-Poulenc Science Book Prize

The winner of this year's £10,000 Rhône-Poulenc Science Book Prize will be announced at the Science Museum in London this week. On the shortlist are:

Climbing Mount Improbable by Richard Dawkins (Penguin, £7.99; pbk). Vigorous popularization of the power and scope of natural selection.

Reviewed in Nature 382, 309 (1996).

Fire in the Mind: Science, Faith and the Search for Order by George Johnson (Penguin, £8.99; pbk). Complexity, Santa Fe-style.

In the Blood: God, Genes and Destiny by Steve Jones (Flamingo, £12.99; pbk). Panoramic view of genetics and human history. Reviewed in Nature 382, 413 (1996).

The Origins of Virtue by Matt Ridley (Viking, £20; hbk). Evolution of altruism and cooperation. Reviewed in Nature 383, 785 (1996).

Longitude by Dava Sobel (Fourth Estate, £12; hbk). How an eighteenth-century clockmaker invented a way of working out location at sea. Reviewed in Nature 385, 309 (1996).

The Wisdom of the Bones: In Search of Human Origins by Alan Walker and Pat Shipman (Phoenix, £7.99; pbk). Husband-and-wife team recount Walker's 1984 discovery of a 1.5-million-year-old Homo erectus skeleton dubbed Nariokotome boy.

Double identity for proteins of the Bcl-2 family

John C. Reed

Bcl-2 is an oncogenic protein that acts by inhibiting programmed cell death. The mechanisms used by this and related anti-apoptotic proteins to protect cells from cytotoxic stimuli are now emerging, with the discovery that Bcl-2 can function both as an ion channel and as an adaptor or docking protein.

Since its discovery over ten years ago as an oncogenic protein involved in human tumours¹, Bcl-2 has defied attempts to determine the biochemical basis for its potent anti-apoptotic action. Recent findings, however, are beginning to reveal details of the mechanisms by which Bcl-2 and its homologous proteins, such as Bcl-X_L, suppress cell death. A picture is emerging of a complicated protein, which in some ways resembles p53 in that it has multiple independent functions, rather like a Swiss army knife.

Although Bcl-2 family proteins lack significant amino-acid sequence homology with other proteins, the three-dimensional structure of Bcl-X_L² reveals a striking similarity to the pore-forming domains of certain bacterial toxins that act as channels for either ions or proteins. The closest in structure to Bcl-X_L are diphtheria toxin, which transports a fragment of the toxin across the cell membrane, and the bacterial pore-forming colicins, which kill *Escherichia coli* by forming non-selective ion channels. If function follows form, the obvious prediction is that Bcl-2 family proteins can function as channels for ions, proteins, or both.

Circumstantial evidence from cell transfection studies indicated that Bcl-2 might have a membrane transport function, with reported effects on Ca²⁺ flux and protein translocation across some of the intracellular membranes where Bcl-2 and its homologues are localized, namely the outer mitochondrial membrane, the endoplasmic reticulum (ER) and the nuclear envelope^{3–5}. Now, direct evidence of ion-channel activity has been obtained from experiments in which the effects of recombinant Bcl-2 or Bcl-X_L in synthetic lipid membranes were studied by using single-channel recordings from planar bilayers and by other approaches^{6,7}. The mechanism by which Bcl-2 and Bcl-X_L create channels in membranes has not been explored in detail, but preliminary indications are that at least some aspects of the process may be similar to the bacterial toxins. For example, the structure of Bcl-X_L reveals a compact seven α -helical bundle, at the centre of which lies a hairpin comprised of two long hydrophobic core α -helices surrounded by five amphipathic α -helices². By analogy to the bacterial toxins, the two core α -helices (α_5 and α_6) presumably insert perpendicularly across the lipid bilayer, with the protein undergoing profound conformational changes, perhaps opening like an umbrella, with the surrounding amphipathic helices folding upwards and resting on top of the membrane. Consistent with this idea, removal of α_5 and α_6 from Bcl-2 prevents channel formation *in vitro* and abolishes its anti-apoptotic effect in cells⁷.

Details about the structures of Bcl-2 and Bcl-X_L channels are still to be resolved and already controversy has developed about the likely size (diameters) of these channels. For example, it has been reported that Bcl-2 and Bcl-X_L form discrete ion channels *in vitro* which mostly assume a closed conformation, open only sporadically, and yield conductances ranging from ~20 to ~300 pS^{6,7}. However, evidence has been found for larger pores under some conditions, which assume a mostly open conformation and produce channels with conductances >1 nS (S. Korsmeyer,

personal communication). Related to the issue of pore diameter is the question of the stoichiometry of these channels. Although the predicted membrane-spanning α_5 and α_6 of Bcl-2 and Bcl-X_L are mostly hydrophobic, as required for insertion into membranes, one face of the helices has hydrophilic residues, which presumably line the aqueous lumen of the channels. By analogy to other natural and synthetic α -helix-type channels (ref. 8), it follows that the membrane-inserted helices from two or more Bcl-2/Bcl-X_L molecules should associate to form a ring that creates an aqueous lumen. How many molecules of Bcl-2 or Bcl-X_L assemble in membranes to create channels is not known, but the existence of multiple Bcl-2/Bcl-X_L conductance states suggests a range of possibilities, at least *in vitro*^{6,7}. There is evidence that Bcl-X_L can function as a monomer, but so far we do not have the methodology to determine the oligomerization state of this protein when it is integrated into membranes and so cannot discount the channel hypothesis. The challenge now is to demonstrate that Bcl-2/Bcl-X_L channels exist *in vivo* and to determine their structures.

Complicating the question of the biological significance of these findings are data indicating that the pro-apoptotic Bcl-2 family member Bax, as well as Bcl-2 and Bcl-X_L, have channel activity *in vitro* (J. C. Martinou, personal communication; S. Schendel, Z. Xie and J.C.R., unpublished results). As Bax can heterodimerize with either Bcl-2 or Bcl-X_L when they are in their compact α -helical-bundle conformations in aqueous environments, these proteins may produce heteromeric channels when integrated into membranes. Alternatively, their heterodimerization may prevent Bcl-2 and Bax from undergoing the conformational changes required for integration into membranes, thus preventing channel formation. Both Bcl-2 and Bax might therefore form cytotoxic channels in cells, with Bcl-2/Bax heterodimerization nullifying channel activity and thus promoting cell survival. This idea predicts that the Bcl-2:Bax ratio is critical, consistent with the finding that excess Bcl-2 paradoxically promotes cell death in some circumstances⁹ and with evidence that Bax and Bak unexpectedly exhibit cytoprotective functions in some cellular contexts^{10,11}. Or Bcl-2/Bcl-X_L and Bax might have different selectivities, providing conduits for different ions or proteins that have opposing effects on apoptosis—this possibility could be tested by mutagenesis in which the presumed channel-forming fifth and sixth helices of these proteins are swapped.

Although at present we have no direct evidence for *in vivo* channel formation, the fact that Bcl-2 family proteins can form channels *in vitro* suggests that they may participate in some of the cellular phenomena recently associated with apoptosis, particularly mitochondrial permeability transition ('megapore' opening) and release of apoptogenic protease activators (cytochrome *c* and apoptosis-inducing factor (AIF)) from mitochondria. Mitochondrial permeability transition involves the opening of a large (~1.3 nS, ~2.9 nm) channel in the inner membrane of the mitochondrion^{12,13}. This occurs almost universally during apoptosis and has repercussions that may contribute to the induction of cell death, including

the generation of oxygen free radicals, dumping of stored Ca^{2+} into the cytosol, and the release of mitochondrial proteins into the cytosol in order to activate the cysteine proteases (caspases) that are the terminal effectors of apoptosis¹⁴. Overexpression of Bcl-2 inhibits mitochondrial permeability transition, whereas Bax overexpression induces it^{15,16}. Bcl-2 is found on the outer mitochondrial membrane and is enriched at the contact sites where the inner and outer membranes abut. As some proteins that regulate the megapore reside in the outer membrane, Bcl-2 and Bax could somehow modulate or participate in megapore formation, possibly by controlling ion fluxes that influence the opening and closing of the megapore. On the other hand, some studies suggest that caspase-activating proteins, in particular cytochrome *c*, can leave the mitochondrial intermembrane space through the outer membrane and enter the cytosol where the cell-death proteases are found, irrespective of mitochondrial permeability transition^{17,18}. If so, channels must exist in the outer mitochondrial membrane that control the release of cytochrome *c* during apoptosis. It remains to be determined whether Bax can create such channels and whether the suppression of cytochrome *c* release by Bcl-2 reflects an ability of Bcl-2 either to plug Bax channels or to transport cytochrome *c* back into mitochondria^{17,18}.

Although the new-found channel activity of Bcl-2 family proteins may prove to be crucial to their function as regulators of cell life and death, these proteins have other important ways of controlling apoptosis. In this regard, Bcl-2 or Bcl-X_L has been reported to bind (or at least coimmunoprecipitate with) eleven proteins, including the protein kinase Raf-1, the protein phosphatase calcineurin, the GTPases R-Ras and H-Ras, the p53-binding protein p53-BP2, the prion protein Pr-1, and several proteins with unknown functions, including CED-4, BAG-1, Nip-1, Nip-2 and Nip-3 (reviewed in ref. 19, and see below). Conversely, the pro-apoptotic protein Bax apparently does not interact with these proteins, suggesting an important difference between the death-suppressing and death-promoting members of the Bcl-2 family.

Although the functional significance of these interactions of proteins with Bcl-2 is unclear, recent discoveries are reducing our ignorance while also raising many questions. The developing theme is that Bcl-2 and its anti-apoptotic homologues, which are normally anchored in intracellular membranes (by virtue of a stretch of hydrophobic residues at their carboxy termini) but oriented towards the cytosol, act at least in part as adaptor or docking proteins which pull other proteins out of the cytosol, either sequestering them to the membrane alongside Bcl-2 (and probably inactivating them) or targeting them for interaction with other membrane-associated proteins.

Of the Bcl-2/Bcl-X_L-binding proteins, CED-4 or its functional equivalents in mammalian cells is probably the most important.

CED-4 is a pro-apoptotic protein from the nematode *Caenorhabditis elegans*²⁰. In a genetic sense, CED-4 functions upstream of the worm homologue of the caspases (CED-3) but downstream of the worm version of Bcl-2 (CED-2) (ref. 21). Although they are functionally connected, it was unclear whether CED-9, CED-4 and CED-3 were physically linked or whether other proteins might be required to create a bridge between them. This speculation was stopped by the recent demonstration that CED-9 as well as Bcl-X_L can bind to worm CED-4, which in turn binds to CED-3 or other caspases²²⁻²⁴. Indeed, CED-9, CED-4 and CED-3 can form a trimolecular complex²³. The mammalian homologue of CED-4 remains at large, but experiments demonstrating co-immunoprecipitation of Bcl-X_L with caspases such as FLICE (caspase-8) and ICE (caspase-1) in human cells indicate that CED-4 homologues may exist in higher organisms. A race is now underway to clone mammalian CED-4 and to define its three-dimensional structure, as well as the mechanism by which it may activate the caspases. In the worm, two forms of CED-4 have been described that arise through alternative splicing of messenger RNA, with the shorter CED-4 promoting apoptosis and the longer CED-4 promoting cell survival²⁵. If this is a sign of things to come in mammalian species, then we can expect an additional level of complexity at the CED-4-dependent step of the cell-death pathway.

Another burning question is how do anti-apoptotic Bcl-2 family proteins such as CED-9 and Bcl-X_L interfere with CED-4 function? Overexpression of worm CED-4 in mammalian cells induces apoptosis in a CED-9- and Bcl-X_L-suppressible manner. CED-4 is normally found in the cytosol, but when co-expressed with CED-9 or Bcl-X_L, it is sequestered to the intracellular membranes where CED-9 and Bcl-X_L are located²⁴. Conversely, when pro-apoptotic members of the Bcl-2 family such as Bax, Bak and Bik are co-expressed with Bcl-X_L, they dimerize with Bcl-X_L, thereby displacing CED-4 (Fig. 1). Thus, it appears that anti-apoptotic Bcl-2 family proteins pull CED-4 out of the cytosol, but the question of exactly how they inhibit CED-4 remains unanswered.

Interactions with other proteins may not be as fundamental to Bcl-2's function as CED-4 binding, but they may account for other activities previously associated with Bcl-2 and provide insight into how Bcl-2 family proteins are regulated. For example, Bcl-2 (but not Bax) can associate with Raf-1, pulling this kinase from the cytosol into proximity with the Bcl-2 on the mitochondrial membrane. Once there, Raf-1 induces phosphorylation of BAD, a pro-apoptotic Bcl-2 family protein that abrogates the cytoprotective functions of Bcl-2 and Bcl-X_L by heterodimerizing with them. The phosphorylated BAD (which unlike Bcl-2 has no membrane-anchoring domain) then dissociates from Bcl-2/Bcl-X_L, moving into the cytosol in a complex with 14-3-3, where it cannot interfere with Bcl-2/Bcl-X_L. Although BAD is one potential substrate of Bcl-2/Raf-1,

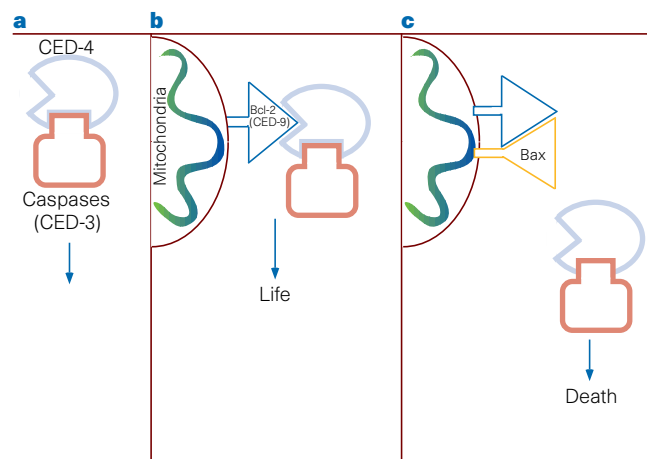
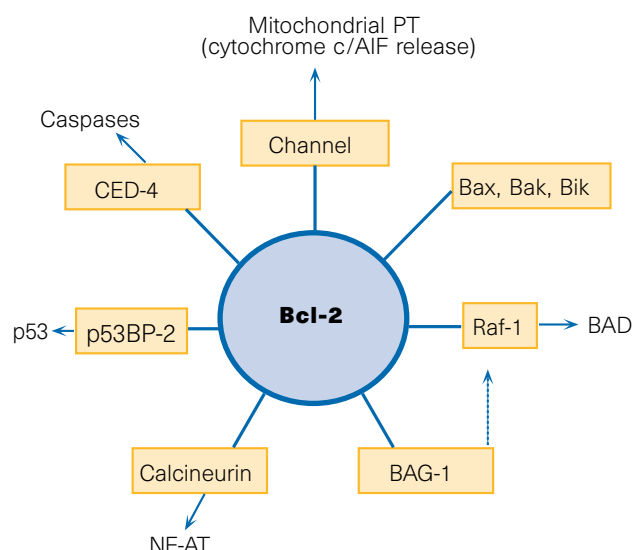


Figure 1 Model for CED-4-mediated bridging of Bcl-2/Bcl-X_L to caspases. **a**, CED-4 binding to cell-death proteases (caspases) such as CED-3 promotes apoptosis. **b**, Bcl-2, tethered to mitochondrial and other intracellular membranes, binds to CED-4, pulling it from the cytosol and presumably preventing activation of caspases. **c**, By heterodimerizing with Bcl-2, pro-apoptotic proteins such as Bax displace CED-4 and presumably allow it to return to the cytosol and to activate proteases.

Functions of Bcl-2 protein



Human Bcl-2 protein

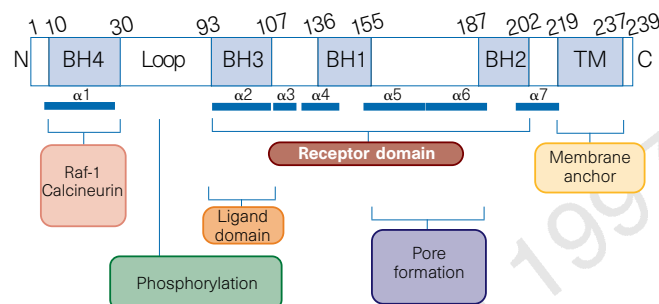


Figure 2 Multifunctional Bcl-2 protein. Left, in addition to channel activity which may directly or indirectly influence mitochondrial megapore opening or release of cytochrome *c* and apoptosis-inducing factor (AIF) from mitochondria, Bcl-2 or Bcl-X_L also binds to several proteins that can participate in cell-death regulation. CED-4 provides a physical connection to the cell-death proteases (caspases). See text for details. Right, the topology of the human Bcl-2 protein is depicted, showing the locations of the Bcl-2 homology (BH) domains, the predicted α -helical segments, and the transmembrane (TM) anchor. Binding of Raf-1 and calcineurin is BH4-

dependent (first α -helix). A long flexible loop between the first and second α -helices is required for Bcl-2 phosphorylation and may represent a negative-regulatory domain. The BH3 domain (second α -helix) plays the role of 'ligand' during dimerization of Bcl-2 family proteins, whereas the combination of the BH1, BH2 and BH3 domains appears to be required for forming the hydrophobic groove into which the BH3 domain inserts. The fifth and sixth α -helices are predicted to participate in channel formation, presumably by penetrating lipid bilayers.

presumably others exist as artificially targeting active Raf-1 to mitochondria can also suppress apoptosis in cells that do not contain BAD protein^{26,27}. Another Bcl-2-binding protein, BAG-1, can also interact with and activate Raf-1. Thus, BAG-1 probably collaborates with Bcl-2, providing a mechanism for activating Raf-1 locally at mitochondrial and other Bcl-2-containing membranes in a Ras-independent manner²⁸.

Little is known about the regions of the Bcl-2 protein needed for interaction with other proteins. However, the three-dimensional structure of Bcl-X_L in its compact, non-membrane-integrated conformation reveals the presence of a hydrophobic groove on the surface of the protein, which is partly formed by the well conserved BH1, BH2 and BH3 domains². The second amphipathic helix of dimerized Bcl-2 family proteins inserts into this groove, like a peptide ligand binding to a receptor^{29,30}. As the BH3 domain corresponds to this inserting α -helix (ligand) but also forms part of the surface pocket (receptor), the implication is that Bcl-2 family proteins assume at least two conformations even in aqueous environments, with one dimerizing partner playing the role of the receptor and the other the ligand. Where CED-4 binds on Bcl-2/Bcl-X_L is unknown, but this surface pocket is a likely candidate, given that some pro-apoptotic Bcl-2 family proteins such as Bik can displace CED-4 from Bcl-X_L and yet have only the BH3 domain (thus limiting their role to that of the ligand). Such 'BH3 only' proteins evidently lack the hydrophobic helices required for membrane integration, suggesting that they can induce apoptosis without forming channels in membranes.

The so-called 'BH4' domain near the N terminus of Bcl-2 is probably not required for binding to CED-4 but is needed for interaction with Raf-1 (refs 25, 26). This domain represents the first α -helix in Bcl-X_L (ref. 2) and is present in all anti-apoptotic Bcl-2 family proteins but absent from nearly all of the pro-apoptotic proteins, implying that Bcl-2/Bcl-X_L evolved an additional domain that provides a means for communicating with Raf-1. The Ca²⁺-dependent protein phosphatase calcineurin also binds to the BH4

domain of Bcl-2 (ref. 31). Overexpression of active calcineurin induces apoptosis in a Bcl-2-suppressible manner³², suggesting a role for this phosphatase during Ca²⁺-mediated apoptosis. Bcl-2 causes a redistribution of calcineurin from the cytosol to intracellular membranes, preventing interaction with phosphorylated NF-AT or other substrates of calcineurin in the cytosol. How calcineurin induces apoptosis is unclear, but dephosphorylation of BAD is one possibility. The Bcl-2/calcineurin interaction may also be relevant to the inhibition by Bcl-2 of cell proliferation and of translocation of NF-AT into the nucleus^{33,34}, given that NF-AT-inducible genes are important for proliferation in some types of cell such as lymphocytes. Although it is not fundamental to its function as a suppressor of apoptosis³⁵, the slowing of entry of quiescent cells into the cell cycle by Bcl-2 could provide additional cytoprotection, because proliferating cells can be more vulnerable to apoptotic stimuli. With kinases and phosphatases hovering around Bcl-2, it is notable that Bcl-2 can be phosphorylated and that this post-translational modification may inactivate Bcl-2 (refs 36, 37) (Fig. 2).

In addition to CED-4, BAG-1, Raf-1 and calcineurin, Bcl-2 can bind to a p53-binding protein, p53-BP2, conceivably explaining how overexpression of Bcl-2 can interfere with translocation of p53 from the cytosol into the nucleus^{5,38}. Immunoelectron microscopy indicates that Bcl-2 is associated with nuclear pore complexes³⁹, an ideal situation for catching proteins as they cross the nuclear envelope.

The available data thus suggest that Bcl-2 has two separate functions as a channel protein and as an adaptor/docking protein (Fig. 2). Although arguments could be raised against both models on the basis that truncation mutants of Bcl-2 without their carboxy-terminal membrane-anchoring domain retain anti-apoptotic function, these mutant proteins still associate with intracellular membranes, probably by virtue of interaction with endogenous Bcl-2 family proteins. The channel and adaptor protein activities may be mutually exclusive in some circumstances, because for channel formation Bcl-2 must integrate into membranes and drastically

alter its conformation, which might prevent it from interacting with CED-4 or other key proteins. Conversely, some protein interactions may be assisted by this integration, for example those of the BH4 domain with Raf-1 and calcineurin. Which of the two functions will prove to be more important? The genetic data from *C. elegans* indicate that CED-4 is likely to be critical for Bcl-2 function, enabling Bcl-2 to regulate the caspases. Moreover, the finding that CED-9 is not required for survival of worms that lack CED-3 or CED-4 argues that channel formation is unnecessary²¹. However, Bax can induce cell death by means that do not depend on caspases, apparently by triggering mitochondrial permeability transition¹⁶.

The dual function of Bcl-2 could explain how Bcl-2 can inhibit both apoptosis and (in some cases) necrosis⁴⁰ by interacting with CED-4 to modulate caspases and regulating mitochondrial permeability transition through its ion-channel activity, respectively. Interaction of Bcl-2 with accessory proteins such as Raf-1, calcineurin, BAG-1 and p53-BP2 will likewise find their niches as additional cell-type- or circumstance-specific embellishments on the theme of Bcl-2/Bcl-X_L as adaptor/docking protein, probably serving as supporting actors in a play in which channel activity and CED-4 binding star as the principal protagonists. □

John C. Reed is at the Burnham Institute, Program on Apoptosis and Cell Death Research, 10901 N. Torrey Pines Road, La Jolla, California 92037, USA.

1. Tsujimoto, Y. & Croce, C. M. Analysis of the structure, transcripts, and protein products of *bcl-2*, the gene involved in human follicular lymphoma. *Proc. Natl Acad. Sci. USA* **83**, 5214–5218 (1986).
2. Muchmore, S. W. *et al.* X-ray and NMR structure of human Bcl-X_L, an inhibitor of programmed cell death. *Nature* **381**, 335–341 (1996).
3. Baffy, G., Miyashita, T., Williamson, J. R. & Reed, J. C. Apoptosis induced by withdrawal of interleukin-3 (IL-3) from an IL-3-dependent hematopoietic cell line is associated with repartitioning of intracellular calcium and is blocked by enforced Bcl-2 oncoprotein production. *J. Biol. Chem.* **268**, 6511–6519 (1993).
4. Lam, M. *et al.* Evidence that Bcl-2 represses apoptosis by regulating endoplasmic reticulum-associated Ca²⁺ fluxes. *Proc. Natl Acad. Sci. USA* **91**, 6569–6573 (1994).
5. Ryan, J. J. *et al.* c-myc and bcl-2 modulate p53 function by altering p53 subcellular trafficking during the cell cycle. *Proc. Natl Acad. Sci. USA* **91**, 5878–5882 (1994).
6. Minn, A. J. *et al.* Bcl-X_L forms an ion channel in synthetic lipid membranes. *Nature* **385**, 353–357 (1997).
7. Schendel, S. L. *et al.* Channel formation by anti-apoptotic protein, Bcl-2. *Proc. Natl Acad. Sci. USA* **94**, 5113–5118 (1997).
8. Montal, M. Protein folds in channel structure. *Curr. Opin. Struct. Biol.* **6**, 499–510 (1996).
9. Chen, J. *et al.* bcl-2 overexpression reduces apoptotic photoreceptor cell death in three different retinal degenerations. *Proc. Natl Acad. Sci. USA* **93**, 7042–7047 (1996).
10. Middleton, G., Nunez, G. & Davies, A. M. Bax promotes neuronal survival and antagonises the survival effects of neurotrophic factors. *Development* **122**, 695–701 (1996).
11. Kiefer, M. C. *et al.* Modulation of apoptosis by the widely distributed Bcl-2 homologue Bak. *Nature* **374**, 736–739 (1995).
12. Bernardi, P., Broekemeier, K. M. & Pfeiffer, D. R. Recent progress on regulation of the mitochondrial permeability transition pore; a cyclosporin-sensitive pore in the inner mitochondrial membrane. *J. Bioenerget. Biomembr.* **26**, 509–517 (1994).

13. Zoratti, M. & Szabo, I. Electrophysiology of the inner mitochondrial membrane. *J. Bioenerget. Biomembr.* **26**, 543–553 (1996).
14. Petit, P. X., Susin, S.-A., Zamzami, N., Mignotte, B. & Kroemer, G. Mitochondria and programmed cell death: back of the future. *FEBS Lett.* **396**, 7–13 (1996).
15. Susin, S. A. *et al.* Bcl-2 inhibits the mitochondrial release of an apoptogenic protease. *J. Exp. Med.* **184**, 1331–1342 (1996).
16. Xiang, J., Chao, D. T. & Korsmeyer, S. J. BAX-induced cell death may not require interleukin 1β-converting enzyme-like proteases. *Proc. Natl Acad. Sci. USA* **93**, 14559–14563 (1996).
17. Kluck, R. M., Bossy-Wetzel, E., Green, D. R. & Newmeyer, D. D. The release of cytochrome c from mitochondria: a primary site for Bcl-2 regulation of apoptosis. *Science* **275**, 1132–1136 (1997).
18. Yang, J. *et al.* Prevention of apoptosis by Bcl-2: release of cytochrome c from mitochondria blocked. *Science* **275**, 1129–1132 (1997).
19. Häcker, G. & Vaux, D. L. A sticky business. *Curr. Biol.* **5**, 622–624 (1995).
20. Yuan, J. Y. & Horvitz, H. R. The *Caenorhabditis elegans* genes *ced-3* and *ced-4* act cell autonomously to cause programmed cell death. *Dev. Biol.* **138**, 33–41 (1990).
21. Shaham, S. & Horvitz, H. R. Developing *Caenorhabditis elegans* neurons may contain both cell-death protective and killer activities. *Genes Dev.* **10**, 578–591 (1996).
22. Spector, M. S., Desnoyers, S., Heppner, D. J. & Hengartner, M. O. Interaction between the *C. elegans* cell-death regulators CED-9 and CED-4. *Nature* **275**, 1122–1126 (1997).
23. Chinnaiyan, A. M., O'Rourke, K., Lane, B. R. & Dixit, V. M. Interaction of CED-4 with CED-3 and CED-9: a molecular framework for cell death. *Science* **275**, 1122–1126 (1997).
24. Wu, D., Wallen, H. D. & Nunez, G. Interaction and regulation of subcellular localization of CED-4 by CED-9. *Science* **275**, 1126–1129 (1997).
25. Shaham, S. & Horvitz, H. R. An alternatively spliced *C. elegans* *ced-4* RNA encodes a novel cell death inhibitor. *Cell* **86**, 201–208 (1996).
26. Wang, H. G., Rapp, U. R. & Reed, J. C. Bcl-2 targets the protein kinase Raf-1 to mitochondria. *Cell* **87**, 629–638 (1996).
27. Zha, J., Harada, H., Yang, E., Jockel, J. & Korsmeyer, S. J. Serine phosphorylation of death agonist BAD in response to survival factor results in binding to 14-3-3 not BCL-X_L. *Cell* **87**, 619–628 (1996).
28. Wang, H.-G., Takayama, S., Rapp, U. R. & Reed, J. C. Bcl-2 interacting protein, BAG-1, binds to and activates the kinase Raf-1. *Proc. Natl Acad. Sci. USA* **93**, 7063–7068 (1996).
29. Sattler, M. *et al.* Structure of Bcl-X_L–Bak peptide complex: recognition between regulators of apoptosis. *Science* **275**, 983–986 (1997).
30. Diaz, J.-L. *et al.* A common binding site mediates heterodimerization and homodimerization of Bcl-2 family members. *J. Biol. Chem.* **272**, 11350–11355 (1997).
31. Shibasaki, F., Kondo, E., Akagi, T. & McKeon, F. Suppression of signalling through NF-AT by interactions between calcineurin and Bcl-2. *Nature* **386**, 728–731 (1997).
32. Shibasaki, F. & McKeon, F. Calcineurin functions in Ca²⁺-activated cell death in mammalian cells. *J. Cell Biol.* **131**, 735–743 (1995).
33. Pietenpol, J. A. *et al.* Paradoxical inhibition of solid tumor cell growth by bcl-2. *Cancer Res.* **54**, 3714–3717 (1994).
34. Linette, G. P., Li, Y., Roth, K. & Korsmeyer, S. J. Cross talk between cell death and cell cycle progression: BCL-2 regulates NFAT-mediated activation. *Proc. Natl Acad. Sci. USA* **93**, 9545–9552 (1996).
35. Huang, D. C. S., O'Reilly, L. A., Strasser, A. & Cory, S. The anti-apoptosis function of Bcl-2 can be genetically separated from its inhibitory effect on cell cycle entry. *EMBO J.* (in the press).
36. Haldar, S., Jena, N. & Croce, C. M. Inactivation of Bcl-2 by phosphorylation. *Proc. Natl Acad. Sci. USA* **92**, 4507–4511 (1995).
37. Chang, B. S., Minn, A. J., Muchmore, S. W., Fesik, S. W. & Thompson, C. B. Identification of a novel regulatory domain in Bcl-X_L and Bcl-2. *EMBO J.* **16**, 968–977 (1997).
38. Naumovski, L. & Cleary, M. L. The p53-binding protein 53BP2 also interacts with Bcl-2 and impedes cell cycle progression at G2/M. *Mol. Cell. Biol.* **16**, 3884–3892 (1996).
39. Krajewski, S. *et al.* Investigations of the subcellular distribution of the bcl-2 oncoprotein: residence in the nuclear envelope, endoplasmic reticulum, and outer mitochondrial membranes. *Cancer Res.* **53**, 4701–4714 (1993).
40. Kane, D. J., Örd, T., Anton, R. & Bredesen, D. E. Expression of Bcl-2 inhibits necrotic neural cell death. *J. Neurosci. Res.* **40**, 269–275 (1995).

Acknowledgements. I thank M. Montal, S. Schendel and S. Choe for discussions on channel proteins, and D. Bredesen and G. Salvesen for critically reviewing the manuscript.

Laser action by tuning the oscillator strength

Jérôme Faist, Federico Capasso, Carlo Sirtori, Deborah L. Sivco, Albert L. Hutchinson & Alfred Y. Cho

Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, Murray Hill, New Jersey 07974, USA

The threshold condition for laser action is usually achieved when the population difference between the initial and final energy levels of the laser transition reaches a critical value, determined by the equality between optical gain and losses. But the threshold condition can also be achieved by increasing the oscillator strength of the laser transition itself, while the population difference is held constant. This forms the basis of a new class of semiconductor lasers, a notable feature of which is broad wavelength tunability (on application of an electric field) in the technologically important mid-infrared region of the spectrum.

A particularly rich area of recent research in semiconductor nanostructures¹ has been that of transitions between sub-bands in quantum wells^{2,3}. These optical transitions connect electronic states within the same band (conduction or valence band) in semiconductor quantum wells, and were first investigated (in the 1960s) in the two-dimensional electron gas at the silicon/silicon dioxide interface⁴. These investigations were subsequently extended to heterojunction quantum wells, thus providing a major impetus to the engineering of new photonic materials and related optoelectronic devices in the mid-infrared region of the spectrum (wavelength 4–12 μm) and beyond^{2,3,5}. Transitions between sub-bands are useful for infrared emitters and detectors^{2,3,6–10} because they make it possible to cover a broad range of wavelengths using the same combination of InP-based and GaAs-based state-of-the-art semiconductor alloys that are used for high-speed electronics and lightwave communications. In these devices one can design and modify at will not only the energy levels and the wavefunctions, the ‘intersubband’ optical matrix elements and the corresponding transition probability, but also the intersubband electron–phonon scattering rates and the electron lifetimes^{3,5,6}. This has led to quantum-cascade lasers in which the intersubband population inversion condition is designed^{6–11}. Inversionless lasers¹² based on intersubband transitions have also been proposed¹³.

Here we report a systematic investigation of a new class of quantum-cascade lasers where quantum design is exploited to

demonstrate a radically new electrical pumping scheme, oscillator strength tuning. In gas and solid-state lasers the threshold condition is achieved by increasing, by either electric or optical pumping, the population difference between the initial and final states of the radiative transition until the material gain equals the losses¹⁴. In the lasers studied here, threshold is instead achieved by keeping the population difference constant and increasing the matrix element and the energy of a suitably designed photon-assisted tunnelling transition. This transition probability tuning is an excellent illustration of how new laser materials can be designed and adds great flexibility to the engineering of semiconductor lasers in the technologically important mid-infrared spectral region.

Laser action by oscillator strength tuning

In our new laser configuration, under application of an electric field of suitable polarity (Fig. 1c), laser action takes place between the lowest energy state ($m = 1$) of one period of thickness L_p and the excited state ($m = 1'$) of the adjacent period downstream. The electron density in the thick trapezoidal well, introduced by doping of the latter, is constant in a wide range of electric fields and temperature owing to the high injection barrier. In particular the ground state has a lifetime which, by design, is much longer (several tens of picoseconds) than that of the other states (≤ 1 ps), including the $m = 1'$ state, in the bias voltage range of the laser and contains the majority of the electron population. The electron density in this

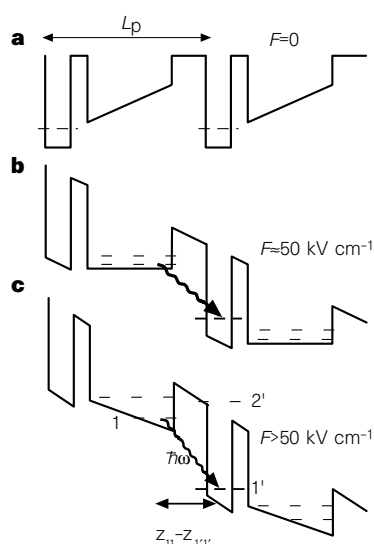


Figure 1 Schematic energy band diagram of two periods of the new laser for different applied electric fields F : **a**, $F = 0$; **b**, $F \approx 50 \text{ kV cm}^{-1}$; **c**, $F > 50 \text{ kV cm}^{-1}$. The electron population is concentrated in the lowest state of the large triangular well created by the applied field. Laser action occurs when the latter is increased until the oscillator strength of the $1-1'$ transition satisfies the threshold condition (gain equals losses; **c**). The wavy arrow indicates the radiative transition. The horizontal dashed lines represent the energy levels and $z_{11} - z_{1'1}$ is the distance between the centroids of the electron probability densities of states 1 and $1'$. L_p is the length of one period. The number of periods is typically 35.

state N_i is therefore also essentially constant in a wide range of currents. Because the ground state $m = 1$ is also the upper state of the lasing transition, the population inversion between this energy level and the empty excited state $1'$ of the adjacent period is maintained regardless of temperature. The laser transition is based on photon-assisted tunnelling^{15–18} ('diagonal' transition) and its energy and matrix element can be strongly tuned by applying an electric field normal to the layers. As the bias is increased, the oscillator strength of the $(1-1')$ transition (which is proportional to the product of the transition energy $E_{1'1}$ and the modulus squared of the transition matrix element) is enhanced. This occurs by two mechanisms: (1) $E_{1'1}$ is shifted to higher values owing to the increased potential drop between the two states (linear Stark

effect); (2) the transition matrix element is increased due to the enhanced spatial overlap (tunnelling) between the states.

When the oscillator strength and the attendant gain cross-section $\sigma_{11'}$ (proportional to the oscillator strength) is such that the gain $\sigma_{11'}N_i$ equals the optical losses, the threshold condition for laser action is achieved. This laser action by oscillator strength tuning is the most fundamental aspect of this new class of lasers and sets it apart from all other mechanisms of achieving threshold so far demonstrated. We have built devices lasing at photon energies corresponding to optical transitions from the ground state (Fig. 1c) for different values of the electric field, thus providing direct evidence of laser action by oscillator strength tuning. In a suitably designed structure we have demonstrated a

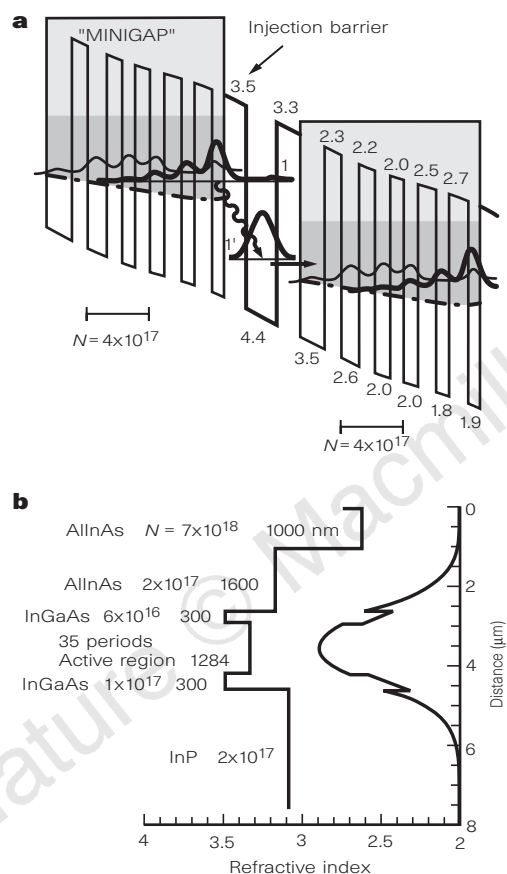


Figure 2 Schematics of the $\text{Al}_{0.48}\text{In}_{0.52}\text{As}/\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ laser heterostructure D-2190. The thicknesses of the layers are given in nm and the doping level is given in cm^{-3} . **a**, Self-consistent calculation of the conduction-band diagram of a portion of the active region at an average electric field of $8.5 \times 10^4 \text{ V cm}^{-1}$. The dot-dashed line is the effective conduction-band edge of the superlattice alloy. The wavy line indicates the transition responsible for laser action. The moduli squared of the relevant envelope wavefunctions are shown. Indicated as 'mini-gap' are the regions with a vanishing density of states: the horizontal segment indicates the doped portion of each period where N is the donor density in cm^{-3} . The material parameters are: conduction band discontinuity $\Delta E_c = 0.52 \text{ eV}$; electron effective mass $m_e^*(\text{Ga}_{0.47}\text{In}_{0.53}\text{As}) = 0.043m_0$, where m_0 is the free electron mass and $m_e^*(\text{Al}_{0.48}\text{In}_{0.52}\text{As}) = 0.078m_0$. Non-parabolicities were taken in account using the method of ref. 20 with the non-parabolicity coefficient $\gamma = 1.13 \times 10^{-18} \text{ m}^2$. **b**, From left to right: composition profile, doping profile, refractive index and mode profile of the waveguide in the direction perpendicular to the layers. The interfaces between the AlInAs and InGaAs regions are graded in composition over a distance of $\sim 30 \text{ nm}$ to smooth out the conduction-band barrier.

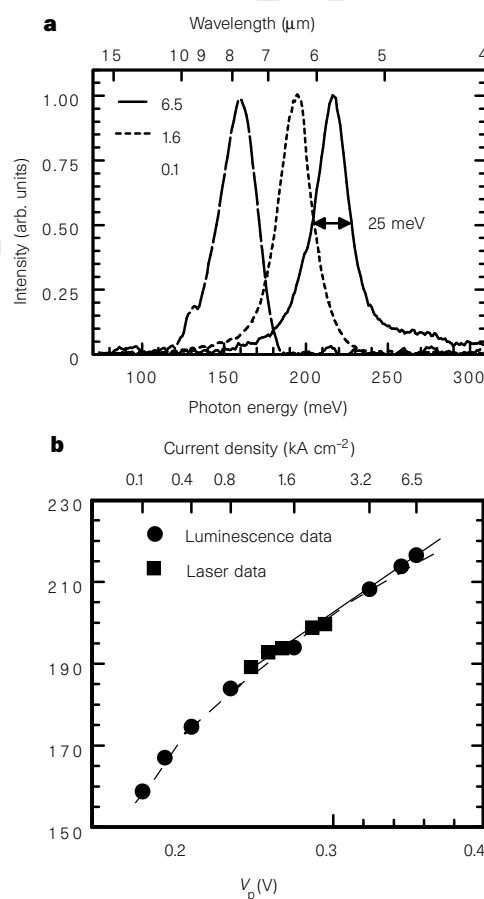


Figure 3 a, Intersubband electroluminescence spectra of sample D-2146 associated with the optical transition from the ground state $m = 1$ to $1'$ (see Fig. 1c) for increasing current densities, shown in key in units of kA cm^{-2} ; solid line, 6.5 kA cm^{-2} ; short-dashed line, 1.6 kA cm^{-2} ; long-dashed line, 0.1 kA cm^{-2} . Sample temperature, 10 K. **b**, Position of the electroluminescence peak as a function of applied voltage per period V_p . The corresponding injected current density is displayed on the top horizontal axis. The straight solid line represents the regime in which the electron density in the $m = 1$ state is constant and equal to the charge introduced by doping. The photon energy is then proportional to the voltage drop between the centroids of the probability densities of states 1 and $1'$ (the linear Stark effect). Sample temperature, 10 K. Filled circles, luminescence sample; filled squares, emission photon energies of five lasers from sample D-2190 with different lengths and widths ranging from 3 to 1.5 mm and from 18 to $10 \mu\text{m}$, respectively, corresponding to different V_p . The coincidence of the laser photon energies with the optical transitions from the $m = 1$ ground state is direct evidence of lasing by oscillator strength tuning. The dashed line is the result of a rigorous calculation including screening effects; the solid line is a linear approximation to the theory (equation (2)) valid when screening is negligible.

large tuning of the wavelength based on the Stark shift of the optical transition.

The present laser differs in significant ways from previous quantum-cascade lasers^{6–10}. In the latter devices, the injection is by resonant tunnelling from the ground state of the injector into an excited state of a coupled or single¹¹ quantum-well active region; the population difference between this energy level and a lower one is increased until threshold is reached. Resonant optical phonon scattering from the final state of the laser transition plays an important role in achieving population inversion in the coupled-well active region structures^{6–10}. In addition, in these quantum-cascade the gain spectrum is not electric-field tunable. Note that in the lasers with single-well active regions, population inversion is difficult to achieve because the upper-state lifetime is very short¹¹.

Laser design and electroluminescence

The transport for increasing applied electric fields can be understood schematically from Fig. 1. At low applied electric fields, the graded region creates a thick effective barrier between the ground states of the wide GaInAs quantum wells (dotted lines in Fig. 1a), and the current flowing through the structure by hopping is negligible. A significant current starts to flow in the structure only for applied fields F larger than $\sim 50 \text{ kV cm}^{-1}$ (corresponding to a bias per period of 0.18 V) necessary to compensate the conduction-band gradient of opposite sign in the trapezoidal wells (Fig. 1b). In the range of fields considered ($F > 50 \text{ kV cm}^{-1}$), the ground state ($m = 1$) of the injector is energetically far away from the ground state $1'$ and the excited state $2'$ of the well in the adjacent period downstream (Fig. 1c). This prevents resonant tunnelling injection into states $1'$ and $2'$ and the attendant negative differential resistance. Transport can then be described in terms of phonon-assisted tunnelling^{15,19} between states 1 and $1'$. This ensures stable injection without high field domains in the whole operating range of our laser, unlike in the resonant tunnelling optical amplifier structures proposed by Kazarinov and Suris¹⁵. In this field range, the compositionally graded region forms a triangular well, the ground state ($m = 1$) of which contains most of the electron population ($>95\%$) at 10 K and is long-lived (tens of picoseconds) compared to other levels. As shown schematically in Fig. 1, and in more detail in Fig. 2a, one period of our structure comprises a compositionally graded well consisting of a short-period $\text{Al}_{0.48}\text{In}_{0.52}\text{As}/\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ superlattice (that is, a pseudo-alloy), and of a $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ quantum well with two $\text{Al}_{0.48}\text{In}_{0.52}\text{As}$ tunnelling barriers. Under applied bias (Fig. 1c), photon-assisted tunnelling emission occurs across the injection barrier upstream from the $\text{Ga}_{0.47}\text{In}_{0.53}\text{As}$ quantum well. The barrier downstream allows fast electron escape into the next well by tunnelling.

The layer sequence of one laser sample (D-2190) is shown in Fig. 2. Thirty-five periods of the active region, grown by molecular

beam epitaxy (MBE) lattice-matched to an InP substrate, are inserted in an optical waveguide similar to those of quantum-cascade lasers^{6–11}. Three other laser samples (D-2152, D-2187, D-2197), based on the scheme shown in Fig. 1 but designed with slightly different parameters, were grown and displayed similar laser performance.

To study the nature of the photon-assisted tunnelling transition ($1-1'$), we investigated intersubband electroluminescence as a function of applied electric field. To suppress optical feedback and gain, we investigated a sample (D-2146) which has the same active region band diagram of the laser sample D-2190, but with a fewer number of periods (five) and without waveguide confining layers. To minimize free-carrier absorption in the substrate, the five periods of the active region—sandwiched between doped ($n^+ = 2 \times 10^{18} \text{ cm}^{-3}$) GaInAs contact layers—were grown on semi-insulating InP substrates. The samples were then processed into 125- μm -diameter circular mesas and the light coupled out through a polished wedge at a 45° angle from the plane of the layers. The electroluminescence spectrum was measured at low temperature with a Fourier-transform infrared spectrometer (FTIR; ref. 17). A few spectra at selected current densities are shown in Fig. 3a.

The large spatial difference $z_{11} - z_{1'1'}$ between the coordinates of the centroids of the electron probability distributions of states 1 and $1'$, and the absence of anticrossing between the $m = 1$ state and any resonance of the adjacent well, cause a strong first-order Stark shift of the $1-1'$ transition. Accordingly, the measured electroluminescence peak shifts from $7.8 \mu\text{m}$ at low current densities (0.1 kA cm^{-2}) to $5.8 \mu\text{m}$ at 6.5 kA cm^{-2} . Diagonal transitions are more sensitive to interface layer roughness than vertical ones; therefore the measured full-width at half-maximum of the luminescence peak ($2\gamma = 25 \text{ meV}$, independent of applied bias) is significantly larger than that observed ($2\gamma = 14 \text{ meV}$) for a vertical transition structure at similar wavelengths^{7,10,11}.

The measured transition energy $E_{1'1}$ of the electroluminescence is plotted as a function of the voltage drop per period V_p in Fig. 3b and compared to the calculated value (dashed line) using a rigorous self-consistent calculation based on Schrödinger's and Poisson's equations²⁰. An example of a calculated energy diagram is shown in Fig. 2a. The agreement is excellent over the whole bias range. The Stark shift of the $1-1'$ transition ($\Delta E_{1'1}$) can be written as

$$\Delta E_{1'1} = q\Delta F[z_{11} - z_{1'1'}] \quad (1)$$

where q is the electron charge and ΔF is increment of the electric field in the region between states 1 and $1'$, neglecting the weak electric-field dependence of $z_{11} - z_{1'1'}$. In the high-field regime, when the electron density is essentially all concentrated in the ground state $m = 1$, for $V_p > 0.22 \text{ V}$, ΔF can be expressed as $\Delta V_p/L_p$ where ΔV_p is the voltage increment across one period.

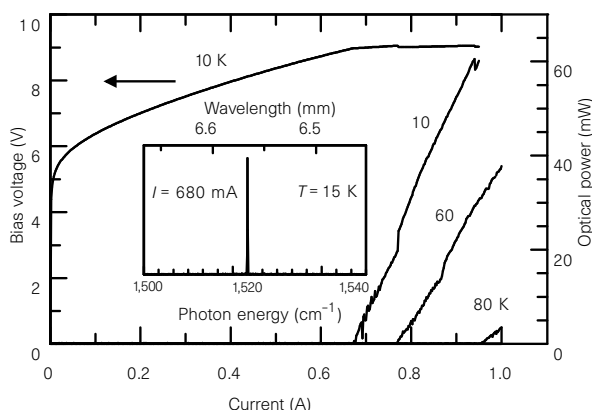


Figure 4 Applied bias and measured continuous-wave optical power from a single facet as a function of injected current for a representative laser of wafer D-2190 at various temperatures. The collection efficiency of the detection apparatus is 50%. The bias voltage is only reported for the lowest temperature (10 K). The device is 3 mm long and $14 \mu\text{m}$ wide. Inset, laser spectrum slightly above threshold ($I = 0.68 \text{ A}$) at $T = 10 \text{ K}$.

Equation (1) then reduces to, in differential form,

$$\frac{1}{q} \frac{dE_{1'1}}{dV_p} \approx \frac{z_{11} - z_{1'1'}}{L_p} \quad (2)$$

Taking $z_{11} - z_{1'1'} = 10$ nm and $L_p = 36.7$ nm, we find a slope $(1/q)(dE_{1'1}/dV_p) = 0.26$ (solid line in Fig. 3) which is in good agreement with the measured value (0.278) obtained from the data of Fig. 3b between $V_p = 0.25$ V and $V_p = 0.32$ V. Note that at a high bias, the theoretical curve slightly deviates from the form given by equation (2). This is due to the small electric-field-induced spatial shifts of states 1 and 1', that is $z_{11} - z_{1'1'}$ decreases with field.

At low bias the experimental Stark shift $(1/q)(dE_{1'1}/dV_p) = 0.54$ is larger than predicted by equation (2). Physically, in this regime, the electron charge progressively transfers from the doped regions of the trapezoidal well into the $m = 1$ state. This transfer tends to screen the field in this region and correspondingly increases the electric field in the undoped region between states 1 and 1' by an amount ΔF greater than $\Delta V_p/L_p$. This screening regime is correctly described by our self-consistent model (dashed curve in Fig. 3b).

In luminescence structures or in lasers below threshold, radiative transitions have a negligible effect on the current density J which can be written as $J = qN_1/\tau_1$ where τ_1 is the lifetime of the $m = 1$ state controlled by optical phonon emission and N_1 is its electron sheet density. In the current density range of our experiments (up to 3 kA cm⁻²), the injected electron density $J\tau_1/q \leq 2 \times 10^{10}$ cm⁻², where J is the measured current density and $\tau_1 \approx 1$ ps is the lifetime

of state $m = 1'$, is small compared to that of the ground state $N_1 = 3.6 \times 10^{11}$ cm⁻², thus preventing space-charge build-up. The absence of space-charge injection in our structure implies global charge neutrality and the same field profile in each period; n_1 is therefore constant and corresponds to the charge introduced by the doping of the injection region ($N_1 = 3.6 \times 10^{11}$ cm⁻²). As the applied field is increased, τ_1 progressively decreases as the wavefunction of state 1 is 'pushed' inside the injection barrier by the progressive narrowing of the triangular well. This variation of τ_1 with constant population N_1 controls the current-voltage characteristic of the device. With $N_1 = 3.6 \times 10^{11}$ cm⁻² (sample D-2146), a lifetime τ_1 of tens of picoseconds (ref. 21) leads to current densities J in the range 1–3 kA cm⁻².

The increase of the penetration of the wavefunction into the injection barrier with increasing field, responsible for the decrease of τ_1 , leads also to a concomitant increase of the optical matrix element $z_{11'}$.

Laser characterization

The lasers were processed into ridge waveguides of width 10–18 μ m at their bases and aligned along the [110] crystallographic direction by wet chemical etching (H₃PO₄:H₂O₂:H₂O, 1:2:4). A 350-nm-thick Si₃N₄ insulation was deposited by chemical vapour deposition. Non-alloyed Ti/Au ohmic contacts were evaporated onto the top layer and the substrate. The lasers were cleaved in bars 1–3.5 mm long and the facets left uncoated. They were then soldered

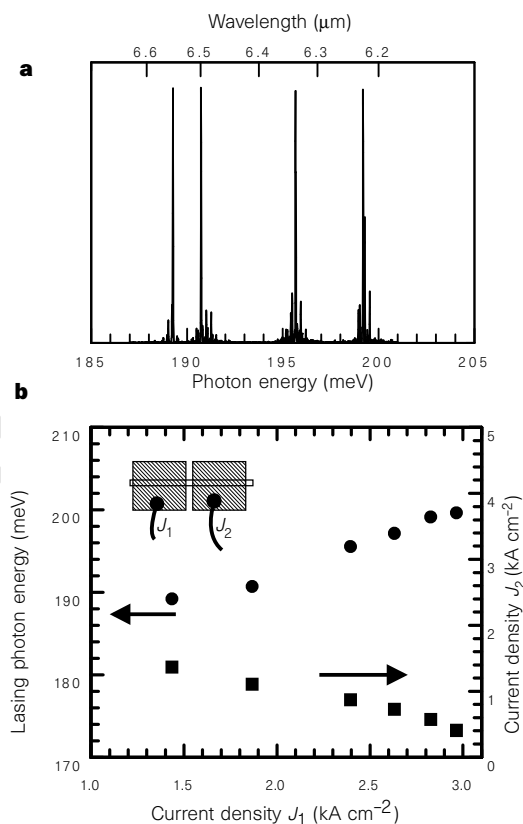


Figure 5 Tunable operation of the laser D-2190. The top metal contact is split into two sections to allow independent electrical injection of two currents densities J_1 and J_2 in the laser waveguide (see inset). The measurements are performed with the laser driven in pulsed mode above threshold with a constant peak optical power of 1 mW. **a**, Measured optical spectra for different values of the currents (J_1 , J_2) from left to right, in units of kA cm⁻²: (1.4,1.4); (1.9,1.1); (2.4,0.9); (2.8,0.6). **b**, Laser photon energy (filled circles) and current density in the second section (filled squares) as a function of injected current density in the first section.

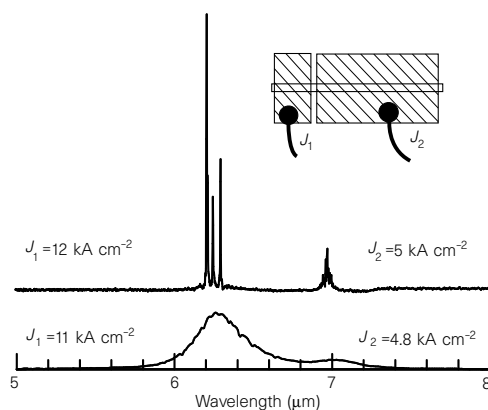


Figure 6 Two-wavelength operation of sample D-2197. The layer sequence of one period of the active region, in nanometres from left to right and starting from the injection barrier (as in Fig. 2a) is **3.5/4.4/3.5/2.8/2.2/2.2/2.0/2.0/2.5/2.1/3.2/2.2**. The barrier layers are in bold; the underlined thickness correspond to the Si-doped layers ($N = 6 \times 10^{17}$ cm⁻³). The length of the small section of the waveguide is 0.8 mm (responsible for the short-wavelength emission) and the long section (responsible for the long-wavelength emission) is 2.8 mm. The lower and upper curves are taken below threshold and at threshold, respectively.

epilayer-up on copper holders, wire-bonded and mounted on a temperature-controlled (10–300 K) cold head of a helium flow cryostat.

Figure 4 shows the continuous-wave optical power versus drive current from a single facet of sample D-2190, obtained using $f/0.8$ optics and a calibrated, room-temperature HgCdTe detector. The laser operates up to 80 K, with up to 40 mW of optical power at 60 K. At 10 K, lasing threshold is obtained at current $I = 0.675$ A and corresponds to a threshold current density $J_{th} = 1.6 \text{ kA cm}^{-2}$. This threshold density is comparable to the best results obtained in previous quantum-cascade lasers with coupled-well active region. The main reason for this is that the decrease of the matrix element of the optical transition and the increase of the transition linewidth is compensated by a much longer electron lifetime.

The voltage applied to this device is displayed in Fig. 4 as a function of injection current at 10 K; the threshold voltage is 9 V. Above threshold, the gain is equal to the total optical losses, pinning the bias voltage at approximately its threshold value, as shown experimentally by the plateau in the voltage–current curve above 0.675 A. Because of this pinning, the laser wavelength does not change significantly as a function of current above threshold. However, because of the strong dependence of the transition energy on the current (Fig. 3), the laser wavelength is expected to vary considerably with threshold current density in devices of different cavity length. More specifically, because the threshold current depends on the cavity loss, the laser photon energy (which is a linear function of voltage) will vary for devices with different lengths and waveguide losses. In Fig. 3, we show (filled squares) the laser photon energy for different values of the threshold bias per period corresponding to different devices with different cavity lengths and widths. The values are in very good agreement with the luminescence measurements. For all the devices, the threshold bias per period is larger than 0.22 V, implying that they operate in the regime of negligible screening. This measurement confirms the peculiar nature of this laser where the threshold condition is met by increasing the optical matrix element with increasing electric field while keeping constant the population of the upper state of the lasing transition. The absence of an independent ‘reservoir’ of electrons might have important implications for the dynamical properties of the laser and for the statistical properties of its stimulated emission.

The spectrum of a device operating in continuous wave above threshold is displayed in Fig. 4 inset. The spectrum is monomode, with a linewidth limited by the resolution of our FTIR spectrometer (0.09 cm^{-1}). Unlike quantum-cascade lasers based on a ‘vertical’ transition²², these devices tend to become multimode when the power exceeds about a milliwatt. This tendency towards multimode operation is consistent with the large non-homogenous broadening of diagonal transitions.

The waveguide loss $\alpha_w = 30 \text{ cm}^{-1}$ was measured by an analysis of the subthreshold luminescence²². The structure is designed to minimize intersubband absorption from the ground state $m = 1$ to the excited states of the same period by avoiding resonance at the laser photon energy. We attribute the waveguide loss to residual intersubband absorption and to free-carrier absorption. At and above threshold, the peak gain equals the losses:

$$\Gamma G_p = \alpha_w - \frac{1}{L} \ln(R) \quad (3)$$

where $\Gamma = 0.52$ is the mode overlap factor. The second term on the right-hand side of equation (3) represents the mirror losses α_m where $R = 0.26$ is the facet reflectivity and $L = 3 \text{ mm}$ is the cavity length of the relatively long device of Fig. 4. The peak material gain G_p is a function of the matrix element $z_{11'}(J)$ of the laser transition⁷

$$G_p = \frac{4\pi q^2 z_{11'}^2(J) N_1}{\epsilon_0 n L_p (2\gamma) \lambda_p} \quad (4)$$

where $\lambda_p = 6.5 \mu\text{m}$ is the peak emission wavelength, $n = 3.2$ is the material’s refractive index and ϵ_0 is the vacuum permittivity. The matrix element $z_{11'}(J)$ increases with current density J , as previously discussed, up to its threshold value $z_{11'}(J_{th}) = 0.39 \text{ nm}$ obtained from equations (3) and (4). The calculated matrix element at the electric field corresponding to the pulsed threshold $J_{th} = 1.4 \text{ kA cm}^{-2}$ is $z_{11'} = 0.41 \text{ nm}$, in good agreement with the previous value.

The laser D-2190 operates in pulsed mode up to temperatures of 220 K with a weak temperature dependence of the threshold current, a characteristic of a laser based on intersubband transitions^{9,10}. Sample D-2197, based on a slightly different active region with a larger number of periods (45), lased close to room temperature (280 K) giving tens of milliwatts of pulsed optical power.

Wavelength tuning and dual wavelength operation

In the structures of our devices, the top contact metallization is not continuous along the whole waveguide but consists of segments separated by $\sim 20 \mu\text{m}$, without etching or cleaving the waveguide, which are contacted in parallel. The resistance between individual sections ($\sim 200 \Omega$) is much larger than the differential resistance of the device, allowing us to inject different current densities in different sections of the device. Let us consider a device in which the waveguide of length $L = 3 \text{ mm}$ is divided in two sections of respective length aL and $(1-a)L$. The specific device considered here has two sections of equal length ($a = 0.5$) (Fig. 5b). In such a structure, keeping the gain constant at its threshold value is only a global requirement, that is the gain in each section (G_1 and G_2) can be independently adjusted as long as the total gain is equal to the total losses

$$\Gamma [a g_1 \mathcal{L}(J_1, \lambda) + (1-a) g_2 \mathcal{L}(J_2, \lambda)] = \alpha_w - \ln(R)/L \quad (5)$$

where $\mathcal{L}(J, \lambda)$ is the lineshape factor and the quantity in parenthesis is the sum of the gains in the two sections. By injecting different current densities J_1 and J_2 in each section, the laser photon energy can now be electrically adjusted (‘tuned’) towards larger energies, approximately tracking the peak of the gain spectrum of the section with the larger injected current as the longitudinal model separation (0.5 cm^{-1}) is negligible compared to the tuning range of the gain spectrum (100 cm^{-1}). In Fig. 5, the measured photon energy and current at threshold in the second section is plotted as a function of the current in the first section. To minimize heating effects, the measurement was performed in pulsed operation. The tuning range shown in Fig. 5 for this laser (6.6 – $6.2 \mu\text{m}$, corresponding to 100 cm^{-1}) is significant for spectroscopic applications.

When the two sections have a significantly different length (that is, for $a < 0.3$ in our case), a striking effect occurs: the structure is able to lase at two wavelengths simultaneously. Having widely different lengths, the two sections will individually reach threshold for a very different current density, and therefore at a very different wavelength because of the Stark tuning. It allows the gain spectrum to display two local maxima which can be brought simultaneously to threshold. We have demonstrated this effect in sample D-2197; Fig. 6 shows sub-threshold luminescence curves for increasing current in the two sections. The two maxima are clearly apparent at wavelengths 7 and $6.2 \mu\text{m}$. Because of gain saturation, we were not able to drive the short section clearly above threshold and the curve at the higher current is taken at threshold.

We have demonstrated a new class of mid-infrared lasers. The importance of the mid-infrared spectrum stems for the existence of atmospheric transmission windows at 3 – $5 \mu\text{m}$ and 8 – $13 \mu\text{m}$. Many gases and vapours have pronounced absorption features in these regions. Thus suitable single-frequency and tunable laser sources in this spectral range can be used for trace-gas analysis and pollution monitoring^{23,24}. The wide tunability of these devices makes them appealing for applications such as spectroscopy of liquids which

have broad absorption lines. Future work will concentrate on increasing the maximum operating temperature and tuning range. The electrical-to-optical power conversion efficiency of these and previous quantum cascade lasers is inherently small (~1%) owing to the high operating currents and voltages. □

Received 8 January; accepted 29 April 1997.

1. Weisbuch, C. & Vinter, B. *Quantum Semiconductor Structures* (Academic, San Diego, 1991).
2. Liu, H. C., Levine, B. F. & Anderson, J. Y. *Quantum Well Intersubband Transitions: Physics and Devices* (NATO ASI Ser. E, Vol. 270, Plenum, New York, 1994).
3. Rosencher, E., Vinter, B. & Levine, B. F. (eds) *Intersubband Transitions in Quantum Wells* (NATO ASI Ser. B, Vol. 288, Plenum, New York, 1992).
4. Ando, T., Fowler, A. B. & Stern, F. Electronic properties of two-dimensional systems. *Rev. Mod. Phys.* **54**, 437–672 (1982).
5. Capasso, F., Faist, J. & Sirtori, C. Mesoscopic phenomena in semiconductor nanostructures by quantum design. *J. Math. Phys.* **37**, 4775–4792 (1996).
6. Faist, J. *et al.* Quantum cascade laser. *Science* **264**, 553–556 (1994).
7. Faist, J. *et al.* Vertical transition quantum cascade laser with Bragg confined excited state. *Appl. Phys. Lett.* **66**, 538–540 (1995).
8. Sirtori, C. *et al.* Long wavelength ($\lambda \sim 11 \mu\text{m}$) quantum cascade lasers. *Appl. Phys. Lett.* **69**, 2810–2812 (1996).
9. Sirtori, C. *et al.* Mid-infrared (8.5 μm) semiconductor laser operating at room temperature. *IEEE Photon. Technol. Lett.* **9**, 294–297 (1997).
10. Faist, J. *et al.* High power mid-infrared ($\lambda \sim 5 \mu\text{m}$) quantum cascade lasers operating above room temperature. *Appl. Phys. Lett.* **68**, 3680–3682 (1996).
11. Faist, J. *et al.* Quantum cascade lasers without intersubband population inversion. *Phys. Rev. Lett.* **76**, 411–414 (1996).

12. Harris, S. E. Lasers without inversion: interference of lifetime-broadened resonances. *Phys. Rev. Lett.* **62**, 1033–1036 (1989).
13. Imamoglu, A. & Ram, R. J. Semiconductor lasers without population inversion. *Opt. Lett.* **19**, 1744–1746 (1994).
14. Sargent, M., Scully, M. O. & Lamb, W. E. *Laser Physics* (Addison Wesley, New York, 1977).
15. Kazarinov, R. F. & Suris, R. A. Electric and electromagnetic properties of semiconductors with a superlattice. *Fiz. Tekh. Poluprov.* **6**, 148–159 (1972); transl. in *Sov. Phys. Semicond.* **6**, 120–131 (1972).
16. Keay, B. J. *et al.* Photon-assisted electric field domains and multiphoton-assisted tunneling in semiconductor superlattices. *Phys. Rev. Lett.* **75**, 4098–4101 (1995).
17. Faist, J. *et al.* Mid-infrared field-tunable intersubband electroluminescence at room temperature by photon-assisted tunneling in coupled-quantum wells. *Appl. Phys. Lett.* **64**, 1144–1146 (1994).
18. Caine, E. J., Subbanna, S., Kroemer, H., Merz, J. L. & Cho, A. Y. Staggered-lineup heterojunctions as sources of tunable below-gap radiation: Experimental verification. *Appl. Phys. Lett.* **45**, 1123–1125 (1984).
19. Oberli, D. Y. *et al.* Direct measurement of resonant and nonresonant tunneling times in asymmetric coupled quantum wells. *Phys. Rev. B* **40**, 3028–3031 (1989).
20. Sirtori, C., Capasso, F., Faist, J. & Scandolo, S. Nonparabolicity and a sum rule associated with bound-to-bound and bound-to-continuum intersubband transitions in quantum wells. *Phys. Rev. B* **50**, 8663–8674 (1994).
21. Price, P. J. Electron transport in polar heterolayers. *Surf. Sci.* **113**, 199–210 (1982).
22. Faist, J. *et al.* Continuous wave operation of quantum cascade lasers based on vertical transitions at $\lambda = 4.6 \mu\text{m}$. *Superlattice Microstructures* **19**, 337–345 (1996).
23. Martinelli, R. U. Mid-infrared wavelengths enhance trace-gas sensing. *Laser Focus World* **32**, 77–81 (1996).
24. Tacke, M. New developments and applications of tunable IR lead salt lasers. *Infrared Phys. Technol.* **36**, 447–463 (1995).

Acknowledgements. We thank S. N. G. Chu for transmission electron microscopy work on some of the structures, and C. Gmachl for discussions.

Correspondence and requests for materials should be addressed to F.C. (e-mail: fc@bell-labs.com).

KNOW YOUR COPY RIGHTS RESPECT OURS

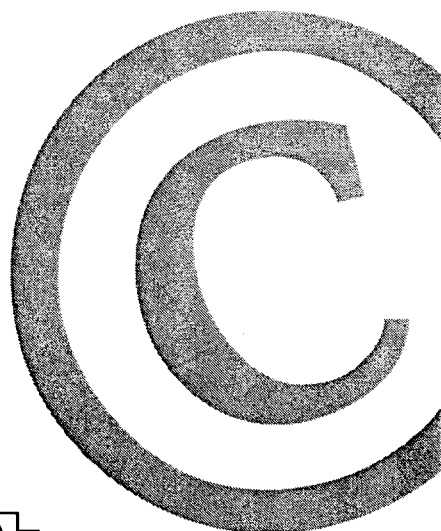
The publication you are reading is protected by copyright law. This means that the publisher could take you and your employer to court and claim heavy legal damages if you make unauthorised photocopies from these pages.

Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

The Copyright Licensing Agency (CLA) is an organisation which issues licences to bring photocopying within the law. It has designed licensing services to cover all kinds of special needs in business, education, and government.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931
Fax: 0171 436 3986

Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997

E. Costa*, F. Frontera†‡, J. Heise§, M. Feroci*, J. in 't Zand§, F. Fiore||, M. N. Cinti*, D. Dal Fiume†, L. Nicastro†, M. Orlandini†, E. Palazzi†, M. Rapisarda#, G. Zavattini‡, R. Jager§, A. Parmar*, A. Owens*, S. Molendi**, G. Cusumano††, M. C. Maccarone††, S. Giarrusso††, A. Coletta‡‡, L. A. Antonelli||, P. Giommil, J. M. Muller||§, L. Piro* & R. C. Butler§§

* Istituto di Astrofisica Spaziale, CNR, Via E. Fermi 21, I-00044 Frascati, Italy

† Istituto Tecnologie e Studio Radiazioni Extraterrestri, CNR, Via Gobetti 101, I-40129 Bologna, Italy

‡ Dipartimento di Fisica, Università Ferrara, Via Paradiso 12, I-44100 Ferrara, Italy

§ Space Research Organization in the Netherlands, Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands

|| Beppo-SAX Scientific Data Center, Via Corcolle 19, I-00131 Roma, Italy

¶ Osservatorio Astronomico di Roma, Monteporzio I-00040 Roma, Italy

Neutronic Section, Fusion Division, ENEA, Via E. Fermi 45, I-00044 Frascati, Italy

†† Astrophysics Division, Space Science Department of ESA, ESTEC, PO Box 299, 2200 AG Noordwijk, Netherlands

** Istituto di Fisica Cosmica e Tecnologie Relative, CNR, Via Bassini 15, I-20133 Milano, Italy

††† Istituto di Fisica Cosmica e Applicazioni Informatica, CNR, Via U. La Malfa 153, I-90139 Palermo, Italy

‡‡ Beppo-SAX Science Operation Center, Via Corcolle 19, I-00131 Roma, Italy

§§ Agenzia Spaziale Italiana, Viale Regina Margherita 202, I-00198 Roma, Italy

Establishing the nature of γ -ray bursts is one of the greatest challenges in high-energy astrophysics. The distribution of these bursts is isotropic across the sky, but inhomogeneous in space, with a deficit of faint bursts¹. It is currently unknown whether γ -ray bursts are produced in our Galaxy or at cosmological distances. The detection and identification of counterparts at other wavelengths are seen as crucial for resolving the origin of the events. Here we report the detection by the Beppo-SAX satellite² of an X-ray 'afterglow', associated with the γ -ray burst of 28 February 1997 (GRB970228; ref. 3)—the first such detection for any γ -ray burst. The X-ray transient was found to contain a significant fraction of the total energy of the γ -ray burst and, following the initial detection⁴ eight hours after the main burst, faded within a few days with a power-law decay function. The rapid locating of this γ -ray burst instigated a multi-wavelength observational campaign that culminated in the identification⁵ of a fading optical transient in a position consistent⁶ with the X-ray transient reported here.

The main reason of our ignorance of the nature of γ -ray burst (GRB) sources is the unfavourable combination of an unusual phenomenology and instrumental inadequacy. γ -ray telescopes have a poor imaging capability and GRBs only last from a fraction of a second to hundreds of seconds. In any case, after a short time they are no longer detectable in the γ -ray band even with large detectors. The burst decay is so fast and the positioning uncertainty so large that no search for delayed emission in other wavelengths has so far been successfully attempted⁷.

The Italian–Dutch Beppo-SAX satellite^{2,8} includes many experiments in different energy bands and with different fields of view. In particular, the combined presence of an all-sky Gamma-Ray Burst Monitor (GRBM)^{9,10} in the 40–700 keV energy range, and two Wide Field Cameras (WFCs)¹¹ in the 2–26 keV energy range, which cover ~5% of the sky with a pixel size of 5 arcmin, allows an unprece-

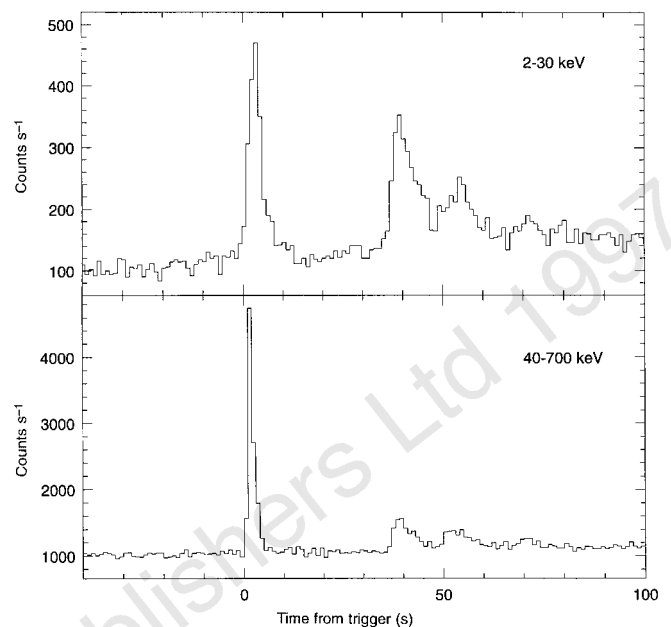


Figure 1 Time profile of GRB970228 in the γ -ray (from the Gamma-Ray Burst Monitor) and X-ray (from the Wide Field Camera) bands. The origin is the trigger time. The first pulse is shorter in γ -rays than in X-rays. Three other pulses follow (at ~35, 50 and 70 s from the trigger) that are much enhanced in the X-ray band. The total burst duration is ~80 s.

dent capability of detecting and fast positioning GRBs and starting follow-up observations.

We developed a procedure for fast localization and rapid follow-up observations of GRBs with the Beppo-SAX Narrow Field Instruments (NFIs), a cluster of telescopes pointing towards the same field of view and covering the large band of 0.1–300 keV (refs 10, 12–14), taking advantage of having them aboard the same satellite and under the same Operation Control Centre.

On 1997 February 28.123620 UT the GRBM was triggered by a GRB event (GRB970228). When the data from the whole orbit were transferred to the ground station and forwarded to the Scientific Operation Centre, 'quick-look' analysis of data from the WFCs at the trigger time showed that a counting excess was also present in one WFC. The X-ray excess was imaged, showing a point-like source. WFC images before and after the event showed that the source was transient and simultaneous with the burst. Light curves in the γ -ray and X-ray band are shown in Fig. 1.

The burst position was first determined from a 'quick-look' analysis of the WFC data with an error radius of ~10 arcmin, suitable for planning a Target of Opportunity pointing (TOO1) of the GRB field with Beppo-SAX NFIs. After few hours, using off-line attitude analysis, we obtained for GRB970228 a refined error box of 3 arcmin radius, centred at right ascension (RA) 05 h 01 min 57 s, declination (dec.) 11° 46.4' (equinox 2000.0). With this refined position, observations in other wavelengths were solicited.

The first observation by the NFIs of Beppo-SAX started on February 28.4681, only 8 hours after the GRBM trigger, and ended on February 28.8330. The total exposure time was 14,344 s in the Medium Energy Concentrator Spectrometer (MECS) and 8,725 s in the Low Energy Concentrator Spectrometer (LECS). In the refined WFC error box we found only one source: 1SAX J0501.7+1146 with coordinates (equinox 2000.0) RA 05 h 01 min 44 s, dec. 11° 46.7' and a 90% confidence error radius of 50 arcsec.

As the pointing of NFIs was based on the first coarse positioning of the GRB in two of the three medium-energy telescopes, the

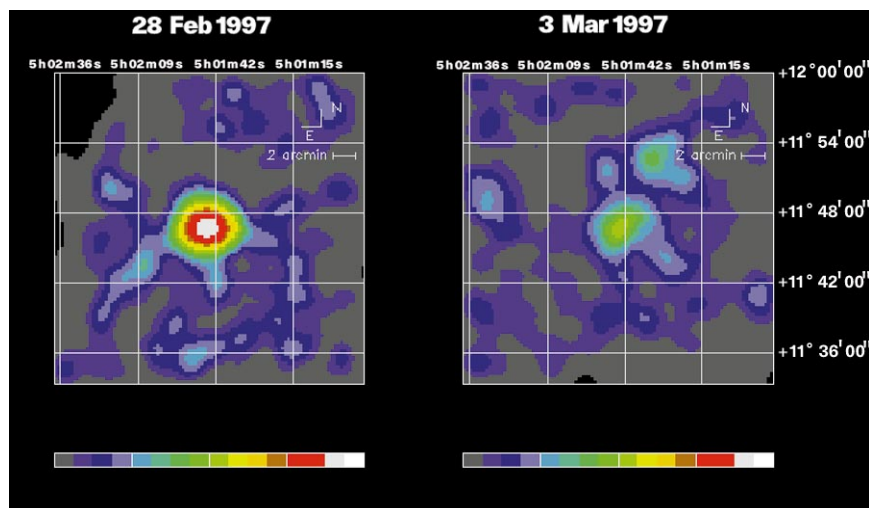


Figure 2 False-colour images of the source 1SAX J0501.7 + 1146, as detected in the error box of GRB970228 with Beppo-SAX Medium Energy Concentrator Spectrometer (2–10 keV) during the first and second Target of Opportunity observations (TOO1 and TOO2, respectively). White corresponds to 31 counts per pixel², green corresponds to 6 counts per pixel² and grey to a background of 0–1 counts per pixel². Taking into account the correction for the number of telescopes (one in TOO1 and three in TOO2) and the vignetting in TOO1 due to off-axis pointing, the source faded by a factor of ~20 in three days. From the ASCA faint sources data³³, the probability that the source detected during the second pointing is coincident by chance with the position of 1SAX J0501.7 + 1146 is of the order of 1×10^{-3} .

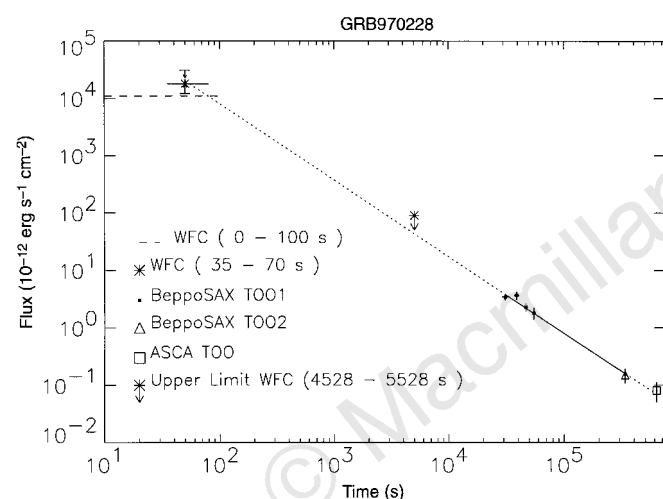


Figure 3 Variation of source flux with time in the 2–10 keV range. Data from the TOO1 observation are grouped into four points of 8,000-s duration each. Data from TOO2 are grouped in one point only due to the lower statistics. The zero time is taken at the GRBM trigger time. Data are fitted by a power law ($\propto t^{-1.32}$). This law is shown as a solid sloping line at lower right. The forward extrapolation of the same law is consistent with the flux detected by ASCA³⁴ on March 7.028 of $(8 \pm 0.3) \times 10^{-14} \text{ erg cm}^{-2} \text{ s}^{-1}$ (averaged value for SIS and GIS detectors), in same energy range. The same law extrapolated backwards (dotted line) to the approximate time of the GRB (described by arrows in the top left) is a good match with the average flux of $2.3 \times 10^{-8} \text{ erg cm}^{-2} \text{ s}^{-1}$ detected by WFC in the three minor pulses of Fig. 1 from 35 to 70 s. Also shown is the 3σ upper limit of the source flux obtained with WFC ~5,000 s after the burst, for an exposure time to the source of 1,000 s.

source was partially covered by the window support structure. To exclude spurious variability due to pointing drifts, in the analysis we only use data from the LECS and only one out of three MECS units.

The source energy spectrum in the 0.1–10 keV band is consistent with a power law of photon index 2.1 ± 0.3 . The hydrogen column density is $(3.5_{-2.3}^{+3.3}) \times 10^{21} \text{ cm}^{-2}$ and consistent with the Galactic absorption along the line of sight $1.6 \times 10^{21} \text{ cm}^{-2}$. The 2–10 keV average source flux during this observation was $(2.8 \pm 0.4) \times 10^{-12} \text{ erg cm}^{-2} \text{ s}^{-1}$, whereas the 0.1–2 keV flux was $(1.0 \pm 0.3) \times 10^{-12} \text{ erg cm}^{-2} \text{ s}^{-1}$ (note that the power-law photon index, the fluxes and the hydrogen column density quoted in ref. 5 are not correct). We also searched for hard X-ray emission (15–100 keV) with the Phoswich Detection System without detecting any line or continuum flux. The 3σ upper limit on the 15–100 keV emission is $4.3 \times 10^{-11} \text{ erg cm}^{-2} \text{ s}^{-1}$, which is higher than the extrapolation from the low-energy power law.

We performed a second Target of Opportunity observation (TOO2) of the field with Beppo-SAX NFIs, about three days after the GRB970228 occurrence time (from March 3.7345 to March 4.1174). The exposure time was 16,270 s with the MECS and 8,510 s with the LECS. A source at a position consistent with that of 1SAX J0501.7 + 1146 was detected in the MECS. Assuming the above spectral shape, the 2–10 keV flux was $(1.5 \pm 0.5) \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$, a factor ~20 lower than in TOO1. The source was not detected in the LECS and the 3σ upper limit in the 0.1–2 keV band was $4 \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$. In Fig. 2 we show the MECS image of the source in the first and in the second observation.

This position is consistent with the GRB error box obtained with WFC, and with the GRB error annulus resulting from the Interplanetary Network (IPN) based on Beppo-SAX GRBM/Ulysses experiments¹⁵.

No source was present in this position in the Rosat all-sky survey¹⁶ with a flux upper limit at 2.5σ of $1.9 \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$, in the range 0.1–2.4 keV, a value compatible with the LECS TOO2 but not with TOO1.

The transient time behaviour and the positional coincidence strongly support the association of 1SAX J0501.7 + 1146 with GRB970228. Using the statistics of X-ray sources derived from the GINGA background analysis¹⁷ we estimate that the probability to have by chance in a field of 3 arcmin radius a source of intensity equal to or higher than the one we detected is $< 8 \times 10^{-4}$. This probability value is reduced by a factor of at least 5 if we take into account the intersection of the error annulus of 30-arcsec half-width derived from IPN for GRB970228^{15,18} with the WFC and NFI error boxes.

Although results of a detailed spectral analysis of 1SAX J0501.7 + 1146 and GRB970228 will be reported elsewhere (F. Frontera *et al.*, manuscript in preparation), we examine here the remarkable time behaviour of the source. Figure 3 shows the 2–10 keV flux evolution during the two TOO observations. The source flux shows a significant decrease within the TOO1 observation. The reduced χ^2 (3 degrees of freedom, d.o.f.) for a constant flux is 3.6, corresponding to a probability of 0.13%. We tried to fit data of both observations with a single law. An exponential decay function does

not fit the data. The best fit of the TOO1 and TOO2 flux data versus time was obtained with a power-law function ($\propto t^{-\alpha}$) (see Fig. 3). The best-fit index is given by $\alpha = 1.33^{+0.13}_{-0.11}$ (χ^2 per d.o.f. = 0.7 with 4 d.o.f.).

We have also compared the flux and the decay law found for 1SAX J0501.7+1146 with the fluxes measured with GRBM and WFC during the γ -ray burst and during the following minor pulses shown in Fig. 1. In Fig. 3 (top left), the dashed line shows the 2–10 keV flux averaged over 100 s corresponding to the entire burst duration, whereas the solid horizontal line gives the average flux of the three minor pulses. Both fluxes are consistent with the backward extrapolation of the derived afterglow decay law. This strongly suggests that the X-ray emission detected soon after the GRB continuously evolves into the X-ray emission of the afterglow.

This result has an implication for the energetics of the event. The GRB fluence measured by GRBM in the 40–700 keV band was $1.1 \times 10^{-5} \text{ erg cm}^{-2}$. The X-ray fluence measured by WFC in the 2–10 keV band was $\sim 1.2 \times 10^{-6} \text{ erg cm}^{-2}$, that is $\sim 11\%$ of the γ -ray fluence. If we assume that the three last pulses in Fig. 1 are part of the afterglow, by integrating the power law from 35 s to infinity we find, in the window 2–10 keV, a fluence which is $\sim 40\%$ of the energy in the γ -ray burst itself in the band 40–700 keV. The X-ray afterglow is not only the low-energy tail of the GRB phenomenon but is also a significant channel of energy dissipation of the event on a completely different timescale.

The well-established power-law decay function of the GRB remnant flux, the consistency of its extrapolation with the X-ray flux at the time of the burst, and the energetic content in X-rays are the main results of our discovery. They will significantly affect models of GRBs and constrain their parameters. Indeed the fast detection of GRB970228, promptly communicated to the scientific community^{3,4}, triggered both the Beppo-SAX NFI follow-up and observations in the radio^{19,20} and optical bands^{21–25}. These observations lead to cogent limits to the radio emission and to the detection^{5,26–30} of an optical transient, in a position consistent with that of 1SAX J0501.7+1146 that faded in a few days. We note, however, that a previous GRB detected by Beppo-SAX, GRB970111³¹, had a γ -ray fluence about four times larger than GRB970228 and an undetectable X-ray emission 16 hours after the burst. No fading optical source was detected at a level of magnitude $B = 23$ and $R = 22.6$ (ref. 32).

The Beppo-SAX measurement, in addition to discovering a relevant delayed X-ray emission, has thus provided the link missing for 25 years between the γ -ray phenomenology and the ultimate location capability of X-ray, optical and radio astronomy. We expect more detections of GRBs by Beppo-SAX GRBM/WFC, along with their follow-up observations. We hope that the existence of X-ray/optical afterglows and their rapid detection will contribute to the unambiguous identification of the GRB sources. $\square$

Received 9 April; accepted 22 May 1997.

1. Fishman, G. J. & Meegan, C. A. Gamma ray bursts. *Annu. Rev. Astron. Astrophys.* **33**, 415–458 (1995).
2. Boella, G. et al. Beppo SAX, the wide band mission for X-ray astronomy. *Astron. Astrophys. Suppl. Ser.* **122**, 299–307 (1997).
3. Costa, E. et al. *IAU Circ.* No. 6572 (1997).
4. Costa, E. et al. *IAU Circ.* No. 6576 (1997).
5. van Paradijs, J. et al. Transient optical emission from the error box of the γ -ray burst of 28 February 1997. *Nature* **386**, 686–688 (1997).
6. Frontera, F. et al. *IAU Circ.* No. 6637 (1997).
7. Vrba, F. Searches for gamma-ray bursts counterparts: current status and future prospects. 565–574 (AIP Proc. 384, Am. Inst. Phys., Huntsville, AL, 1996).
8. Piro, L., Scarsi, L. & Butler, R. C. in *X-Ray and EUV/FUV Spectroscopy and Polarimetry* (ed. Fenischi, S.) 169–181 (SPIE 2517, San Diego, 1995).
9. Costa, E. et al. The gamma-ray burst monitor onboard SAX. *Adv. Space Res.* (submitted).
10. Frontera, F. et al. The high energy instruments PDS on-board the BeppoSAX X-ray astronomy satellite. *Astron. Astrophys. Suppl. Ser.* **122**, 357–369 (1997).
11. Jager, R. et al. The Wide Field cameras onboard the BeppoSAX X-ray astronomy satellite. *Astron. Astrophys. Suppl. Ser.* (submitted).
12. Parmar, A. et al. The low energy concentrator spectrometer on-board the BeppoSAX X-ray astronomy satellite. *Astron. Astrophys. Suppl. Ser.* **122**, 309–326 (1997).
13. Boella, G. et al. The medium energy concentrator spectrometer on-board the BeppoSAX X-ray astronomy satellite. *Astron. Astrophys. Suppl. Ser.* **122**, 327–340 (1997).
14. Manzo, G. et al. The high pressure gas scintillation proportional counter on-board the BeppoSAX X-ray astronomy satellite. *Astron. Astrophys. Suppl. Ser.* **122**, 341–356 (1997).

15. Hurley, K. et al. 3rd Interplanetary Network localization, time history, fluence, and peak flux of the February 28, 1997 gamma-ray burst. *Astrophys. J. Lett.* (in the press).
16. Boller, T. et al. *IAU Circ.* No. 6580 (1997).
17. Warwick, R. & Stewart, G. in *Proc. 23rd ESLAB Symp. on Two-Topics in X-Ray Astronomy* (eds Hunt, J. & Batteic, B.) 727–731 (SP-296, ESA, Bologna, 1989).
18. Cline, T. L. et al. *IAU Circ.* No. 6593 (1997).
19. Galama, T. J. et al. *IAU Circ.* No. 6574 (1997).
20. Frail, D. A. et al. *IAU Circ.* No. 6576 (1997).
21. Metzger, R. M. et al. *IAU Circ.* No. 6588 (1997).
22. Pedersen, H. et al. *IAU Circ.* No. 6580 (1997).
23. Wagner, et al. *IAU Circ.* No. 6581 (1997).
24. Tonry, J. L. et al. *IAU Circ.* No. 6620 (1997).
25. Metzger, M. R. et al. *IAU Circ.* No. 6631 (1997).
26. Guarnieri, A. et al. Earliest detection of an optical transient following the gamma-ray burst GRB970228. *Astron. Astrophys. Lett.* (submitted).
27. Margon, B. et al. *IAU Circ.* No. 6618 (1997).
28. Pedichini, F. et al. *IAU Circ.* No. 6635 (1997).
29. Sahu, K. et al. *IAU Circ.* No. 6606 (1997).
30. Sahu, K. et al. *IAU Circ.* No. 6619 (1997).
31. Costa, E. et al. *IAU Circ.* No. 6533 (1997).
32. Castro-Tirado, A. et al. *IAU Circ.* No. 6598 (1997).
33. Georgantopoulos, I. et al. ASCA observations of deep ROSAT fields I: the nature of the X-ray source populations. *Mon. Nat. R. Astron. Soc.* (in the press).
34. Yoshida, A. et al. *IAU Circ.* No. 6593 (1997).

Acknowledgements. This research is supported by the Italian Space Agency (ASI) and Consiglio Nazionale Ricerche of Italy. Beppo-SAX is a joint program of ASI and the Netherlands Agency for Aerospace Programs (NIVR). We thank the staff at the Beppo-SAX Scientific Operation Centre and the Operation Control Centre for their contributions to the GRB research program.

Correspondence should be addressed to E.C. (e-mail: costa@saturn.ias.fra.cnr.it).

A possible long-lived belt of objects between Uranus and Neptune

Matthew J. Holman

Canadian Institute for Theoretical Astrophysics, McLennan Physical Laboratories, University of Toronto, 60 St George Street, Toronto, Ontario M5S 3H8, Canada

Recent discoveries of objects orbiting beyond Neptune^{1–5} have emphasized that our understanding of the distribution and dynamics of material in the outer Solar System is very incomplete. This trans-neptunian population—known as the Kuiper belt—is thought to act as a relatively stable reservoir of objects that could become short-period comets^{6–9}, although there may be other regions of stability in the outer Solar System that could also supply such comets. Here I use numerical simulations to identify one such long-lived region between the orbits of Uranus and Neptune. I show that in the region 24–27 AU from the Sun, about 0.3 per cent of an initial population of small bodies moving on low-eccentricity, low-inclination orbits could survive for the age of the Solar System. The actual existence of this hypothetical belt is not precluded by currently available observational limits, and there could be as much as $\sim 5 \times 10^{-4}$ Earth masses of material populating this region—comparable to the mass of the asteroid belt between Mars and Jupiter.

Numerical investigations show there are few such regions in the outer Solar System where material could survive for the age of Solar System without being ejected through the gravitational influence of the giant planets^{7,10}. In a recent study⁷, a few thousand test particles, starting from circular orbits in the plane of the Solar System from 5 to 50 AU from the Sun, were numerically integrated for times up to 800 Myr. Later work extended this to 4.5 Gyr for the regions between the outer planets and to 1 Gyr beyond Neptune¹¹. During the integrations, any test particle that entered the gravitational sphere of influence of a planet¹² was removed from the simulation. Such test particles are typically ejected from the Solar System in another 10^6 – 10^7 yr (ref. 13). The results are shown in Fig. 1. The green points mark the close encounter times as a function of semimajor axis; the red points indicate those test particles that survived the full

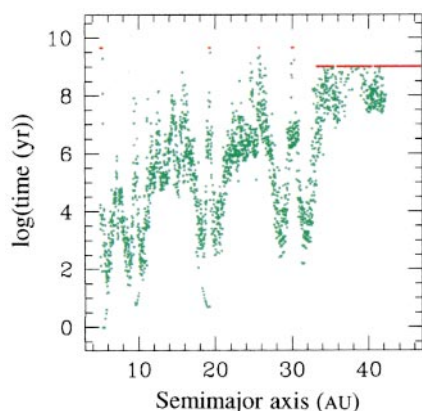


Figure 1 Time survived by each test particle as a function of initial semimajor axis. Green points, times at which the test particles encountered a planet; red points, those test particles which survived the full integration, 4.5 Gyr interior to Neptune and 1 Gyr exterior to Neptune. Beyond ~ 43 AU all the test particles survive the full integration. The scatter of points gives an idea of the spread of termination times for any given semimajor axis. The spikes at 5.2, 9.5, 19.2 and 30.1 AU, at the semimajor axes of the outer planets, correspond to test particles librating in Trojan or horseshoe-like orbits. A few test particles remain at locations of Jupiter, Uranus and Neptune. The absence of Saturn Trojans may not be significant as the test particles were not explicitly placed near the L_4 and L_5 Lagrange points of the outer planets. Between the planets nearly all of the test particles have been removed.

integration. Beyond Neptune (at 30.1 AU), the gravitational erosion in the Kuiper belt is apparent, yet regions stable on 10^9 -yr timescales can be clearly identified. Apart from a few test particles near the L_4 and L_5 Lagrange points of the outer planets, nearly all of the test particles started in the outer planet region have been removed. Here I note that one solitary particle, at 25.7 AU, remains in the 24–27 AU region after 4.5 Gyr.

To determine the significance of this single survivor, I integrated the orbits of several thousand additional test particles in the 24–27 AU region. The initial semimajor axes of the test particles were distributed uniformly from 24 to 27 AU. The initial eccentricity and inclination ranges were $e_0 = 0.00$ – 0.05 and $i_0 = 0^\circ$ – 10° , respectively. The orbits were integrated for times up to 4.5 Gyr. No test particle with $e_0 > 0.03$ or $i_0 > 3^\circ$ survived for more than 1 Gyr; no test particle with $e_0 > 0.01$ or $i_0 > 1^\circ$ survived for the full 4.5 Gyr. Including the single survivor from the initial simulation, five test particles out of 1,700 (or 0.3%) with $e_0 \leq 0.01$ and $i_0 \leq 1^\circ$ survived 4.5 Gyr. Figure 2 shows the time survived as a function of initial semimajor axis for those test particles with $e_0 \leq 0.01$ or $i_0 \leq 1^\circ$ in the new simulation. Two regions, near 24.6 AU and 25.6 AU, harbour the longest-lived test particles.

Figure 3 shows the number of test particles remaining in the 24–27 AU region as a function of time. Assuming a power-law decay profile after 0.1 Gyr, a maximum-likelihood analysis yields an exponent of -1.01 ± 0.13 , in agreement with the t^{-1} decay (where t is the elapsed integration time) of objects with planet-crossing orbits¹⁴. The dashed line, with slope of -1 , is given for reference. Extrapolating from 236 test particles remaining at 0.1 Gyr we would expect 5 ± 2 test particles at 4.5 Gyr, assuming t^{-1} decay. The remaining test particles are not anomalies but the end result of the slow decay of an initially larger population. The survivors are not stable in the sense that they will remain indefinitely: their orbits are chaotic with Lyapunov times of $\sim 6 \times 10^5$ yr. Probably, the t^{-1} decay profile will continue until all the test particles are removed.

Other processes may have significantly reduced the population of bodies between Uranus and Neptune. Collisions are thought to have eroded the population in the inner regions of the Kuiper belt^{15,16} and

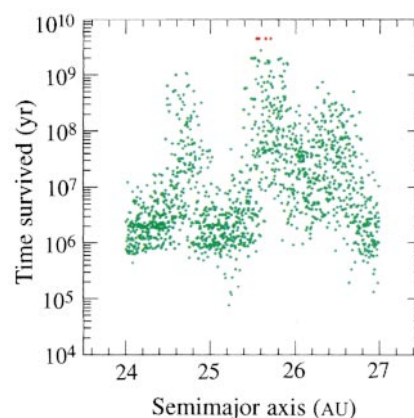


Figure 2 Results of additional simulations of the region between Uranus and Neptune. The test particles here had low initial eccentricity and inclination, $e_0 \leq 0.01$ and $i_0 \leq 1^\circ$. Again, the green points mark the times at which test particles encountered a planet. The red points indicate those test particles which survived the full 4.5-Gyr integration. The regions near 24.6 AU and 25.6 AU harbour the longest-lived members of the 24–27 AU region.

could have had a similar effect in the Uranus–Neptune region. Likewise, planet migration or the presence of large scattering bodies during the late stages of planet formation would tend to increase the typical eccentricity or inclination of the population, thereby reducing the fraction of material surviving to the present^{17–20}. However, if we assume that that collisional processes removed a similar fraction of material from the Uranus–Neptune region as from the Kuiper belt, and if we assume that primordial material had low eccentricities and inclinations, then we can compare the number of objects in the Uranus–Neptune region to that in the Kuiper belt. The current number of bodies indigenous to the 24–27 AU belt is roughly

$$N_{\text{ind}} \propto \sigma_{25} A_{\text{ind}} f_{\text{ind}} \quad (1)$$

where σ_{25} is the initial mass surface density at 25 AU, A_{ind} is the area of the 24–27 AU belt, and f_{ind} is the fraction of the population not removed by the gravitational influence of the giant planets. From the numerical integrations, $f_{\text{ind}} \approx 0.003$. The number of Kuiper-belt objects around 40 AU is, correspondingly,

$$N_k \propto \sigma_{40} A_k f_k \quad (2)$$

Here I adopt $f_k \approx 0.4$ (refs 7, 21). Assuming $\sigma(r) \approx r^{-2}$ (ref. 22) (where r is the heliocentric distance), and taking the 24–27 AU belt and the Kuiper belt to be annuli, 3 AU and 10 AU wide, centred on 25 AU and 40 AU, respectively, then

$$N_{\text{ind}} \approx \left(\frac{40}{25}\right)^2 \left(\frac{25 \times 3}{40 \times 10}\right) \left(\frac{0.003}{0.4}\right) N_k = 4 \times 10^{-3} N_k \quad (3)$$

There are estimated⁴ to be 7×10^4 objects in the Kuiper belt with radii greater than 50 km. Equation (3) implies that there are ~ 300 objects of similar size in the 24–27 AU belt.

This estimated population in the region does not violate available observations, despite assumptions that tend towards an overestimate. Two surveys in particular constrain the population in the 24–27 AU region. The first survey, by Kowal²³, examined 6,400 square degrees of sky to an R-band limiting magnitude $m_R = 18.5$. The limiting magnitude has been adjusted for an estimated trailing loss

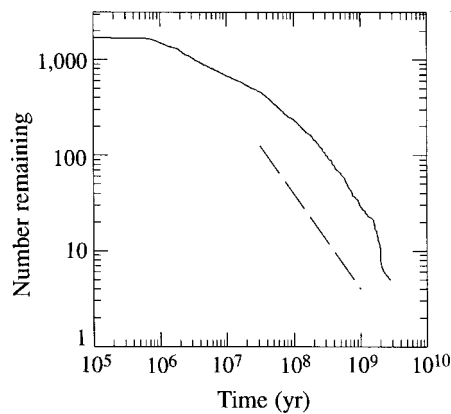


Figure 3 The solid line marks the number of test particles remaining as a function of integration time, for those particles started in the interval 24–27 AU. After 0.1 Gyr the profile approaches t^{-1} decay. The dashed line, with a slope of -1 , is shown for reference.

of ~ 1 magnitude (ref. 4). Kowal's survey yielded 2060 Chiron, the first Centaur comet. The second survey, by Jewitt, Luu and Chen, combines two fainter searches of 1.2 and 3.9 square degrees near the ecliptic to $m_R = 24.2$ that together yielded discoveries of 17 Kuiper-belt objects and two Centaurs (refs 1, 4). The Centaurs discovered to date have semimajor axes outside of 24–27 AU and/or high eccentricities: thus they are not indigenous members of the 24–27 AU region. The observations can only provide an upper limit to the number of objects that could occupy this region and still reasonably escape detection. The apparent magnitude of an object is related to its absolute magnitude by

$$m_R = H_R + 5 \log_{10}(r\Delta) \quad (4)$$

where r is the heliocentric distance (in AU) and $\Delta = r - 1$ is the geocentric distance at opposition (also in AU). For objects with $r = 25$ AU, the searches could detect objects with $H_R < 4.6$ and $H_R < 10.3$, respectively. Neglecting phase-angle effects, the radius, z , of an object (in kilometres) given its absolute magnitude H_R is

$$z = 3.1 \times 10^3 \left(\frac{0.03}{p_R} \right)^{1/2} 10^{-0.2H_R} \quad (5)$$

where p_R is its R-band geometric albedo (ref. 2). Assuming $p_R = 0.04$, $r = 25$ AU, and $\Delta = 24$ AU, Kowal's survey and the Jewitt–Luu–Chen survey could only detect objects with radii $z > 320$ km and $z > 25$ km, respectively. At 99.9% confidence there can be at most seven objects with $z > 320$ km, and 2,900 objects with $z > 25$ km in a 6° wide ecliptic band. This assumes all objects with inclinations greater than 3° have been removed dynamically.

The relative importance of these two constraints depends on the size distribution in the region. The number of objects with radii greater than z can be approximated by a power law such that

$$N(>z) \propto z^{1-q} \quad (6)$$

where $3 \leq q \leq 5$ (from studies of Kuiper-belt objects and short-period comets^{2,4,9,24}). For $q = 3$, Kowal's survey imposes the stronger constraint, $N(> 50 \text{ km}) < 300$. For $q = 5$, the Jewitt–Luu–Chen survey is more restrictive, $N(> 50 \text{ km}) < 180$. Both constraints are comparable to the estimate of $N(> 50 \text{ km}) \approx 300$ from equation (3). Assuming an average density of $\rho = 1,000 \text{ kg m}^{-3}$ and arbitrarily considering only objects with radii between 10 and 500 km, the observational upper limits to the mass in the 24–27 AU region are $3 \times 10^{21} \text{ kg}$ (5×10^{-4} Earth masses) for $q = 3$ and $2 \times 10^{21} \text{ kg}$ (3×10^{-4} Earth masses) for $q = 5$. For comparison, the mass of the asteroid belt is estimated^{25,26} to be $3 \times 10^{21} \text{ kg}$.

Figures 1–3 illustrate that, if the 24–27 AU belt is shown to be populated, material would still be escaping from this region. Just as the Kuiper belt serves as a reservoir and source of material, the bodies in the 24–27 AU region that develop Uranus- or Neptune-crossing orbits could contribute to the flux of ecliptic comets in the outer planet region. The steady-state population of such comets originating in the 24–27 AU belt is roughly

$$N_{c,\text{ind}} \approx \frac{\tau_c}{\tau_{\text{ind}}} N_{\text{ind}} \quad (7)$$

where τ_c is the dynamical lifetime of a comet leaving the 24–27 AU region and τ_{ind} is the dynamical lifetime of a body in the 24–27 AU region before it develops a planet-crossing orbit. If we assume that the lifetime of a comet originating in the 24–27 AU belt is similar to that of a comet from the Kuiper belt, then⁹ $\tau_c = 4.5 \times 10^7 \text{ yr}$. From the t^{-1} decay of the population in the 24–26 AU belt, we get $\tau_{\text{ind}} \approx 4.5 \times 10^9 \text{ yr}$. These values yield

$$N_{c,\text{ind}} \approx 10^{-2} N_{\text{ind}} = 4 \times 10^{-5} N_k \quad (8)$$

For comparison the steady-state number of ecliptic comets originating in the Kuiper belt is roughly

$$N_{c,k} \approx \frac{\tau_c}{\tau_k} N_k = 2 \times 10^{-3} N_k \quad (9)$$

where $\tau_k = 2 \times 10^{10} \text{ yr}$ is the typical dynamical lifetime of a Kuiper-belt object before it encounters Neptune and $\tau_c = 4.5 \times 10^7 \text{ yr}$ as before^{9,21}. By this comparison, the 24–27 AU belt adds 2% to the steady-state number of Centaurs and short-period comets in the outer-planet region. Caution should be exercised here as all the numbers used are uncertain.

The simple calculations presented here open a number of avenues for future research. First, other dynamical processes such as collisions or outer-planet migration^{17–20} could be explored to assess their effect on the distribution of surviving objects in the 24–27 AU region. Second, the L_4 and L_5 points of the outer planets, as well as other regions that could potentially harbour residual mass, should be carefully considered. Third, surveys with fainter limiting magnitudes and more extensive sky-plane coverage will either reveal members of the 24–27 AU belt or place more significant limits on the population in the region. A survey of 8 square degrees to $m_R = 26.2$, sensitive to objects with radii $z > 10 \text{ km}$ at 25 AU assuming $p_R = 0.04$, would reduce the current upper limit on $N(> 10 \text{ km})$ by an order of magnitude. Although such a survey would be a tremendous undertaking with today's technology, it could be more easily accomplished by using telescopes with larger collecting areas and larger fields of view; such telescopes should soon become available. The discovery of material in the 24–27 AU region would imply either that the processes occurring during planet formation left the material in this region undisturbed or that some other process later repopulated the region. The absence of members of the 24–27 AU belt would imply the occurrence of some event, such as planet migration, that disrupted this fragile population. $\square$

Received 2 December 1996; accepted 16 April 1997.

- Jewitt, D. & Luu, J. X. Discovery of the candidate Kuiper belt object 1992QB₁. *Nature* **362**, 730–732 (1993).
- Irwin, M., Tremaine, S. & Zytlow, A. N. A search for slow-moving objects and the luminosity function of the Kuiper belt. *Astron. J.* **110**, 3082–3092 (1995).
- Williams, I. P., O'Ceallaigh, D. P., Fitzsimmons, A. & Marsden, B. G. The slow-moving objects 1993SB and 1993SC. *Icarus* **116**, 180–185 (1995).
- Jewitt, D., Luu, J. & Chen, J. The Mauna Kea-Cerro-Tololo (MKCT) Kuiper belt and Centaur survey. *Astron. J.* **112**, 1225–1238 (1996).
- Cochran, A. L., Levison, H. F., Stern, S. A. & Duncan, M. J. The discovery of Halley-sized Kuiper belt objects using the *Hubble Space Telescope*. *Astrophys. J.* **455**, 342–346 (1995).
- Duncan, M., Quinn, T. & Tremaine, S. *Astrophys. J.* **328**, L69–L73 (1988).
- Holman, M. J. & Wisdom, J. Dynamical stability in the outer solar system and the delivery of short period comets. *Astron. J.* **105**, 1987–1999 (1993).
- Levison, H. F. & Duncan, M. J. The gravitational sculpting of the Kuiper belt. *Astrophys. J.* **406**, L35–L38 (1993).
- Levison, H. F. & Duncan, M. J. From the Kuiper belt to Jupiter-family comets: the spatial distribution of ecliptic comets. *Icarus* (in the press).
- Gladman, B. & Duncan, M. J. On the fates of minor bodies in the outer solar system. *Astron. J.* **100**, 1680–1693 (1990).

11. Holman, M. in *Proc. 27th Symp. on Celestial Mechanics* (eds Kinoshita, H. & Nakai, H.) 116–136 (Nat. Astron. Observatory of Japan, Tokyo, 1995).
12. Roy, A. E. *Orbital Motion* (Hilger, Bristol, 1988).
13. Levison, H. F. & Duncan, M. J. The long-term dynamical behavior of short-period comets. *Icarus* **108**, 18–36 (1994).
14. Dones, L., Levison, H. F. & Duncan, M. *Completing the Inventory of the Solar System* (eds Rettig, T. W. & Hahn, J. M.) 233–244 (ASP Conf. Proc., Astron. Soc. Pacif., San Francisco, 1996).
15. Stern, S. A. Collision timescales in the Kuiper disk: model estimates and their implications. *Astron. J.* **110**, 856–868 (1995).
16. Davis, D. R. & Farinella, P. Collisional evolution of Edgeworth-Kuiper belt objects. *Icarus* **125**, 50–60 (1997).
17. Fernandez, J. A. & Ip, W. H. Some dynamical aspects of the accretion of Uranus and Neptune: the exchange of orbital angular momentum with planetesimals. *Icarus* **58**, 109–120 (1984).
18. Malhotra, R. The origin of Pluto's peculiar orbit. *Nature* **365**, 819–821 (1993).
19. Malhotra, R. The origin of Pluto's orbit: implications for the solar system beyond Neptune. *Astron. J.* **110**, 420–429 (1995).
20. Morbidelli, A. & Valsecchi, G. B. Neptune scattered planetesimals could have sculpted the primordial Edgeworth-Kuiper belt. *Icarus* (in the press).
21. Duncan, M. J., Levison, H. F. & Budd, S. M. The dynamical structure of the Kuiper belt. *Astron. J.* **110**, 3073–3081 (1996).
22. Tremaine, S. in *Baryonic Dark Matter* (eds Lynden-Bell, D. & Gilmore, G.) 37–65 (Kluwer, Boston, 1990).
23. Kowal, C. T. A solar system survey. *Icarus* **77**, 118–123 (1989).
24. Weissman, P. R. & Levison, H. F. The size distribution of cometary nuclei. *Lunar Planet. Sci.* **27**, 1409 (1996).
25. Kresak, L. Mass content and mass distribution of the asteroid system. *Bull. Astron. Inst. Czech.* **28**, 65–82 (1977).
26. Morrison, D. Asteroid sizes and albedos. *Icarus* **31**, 185–220 (1977).

Acknowledgements. I thank N. Murray and S. Tremaine for discussions, S. Dermott and H. Levison for reviews of the manuscript, and J. Wisdom for computer time for the simulations.

Correspondence should be addressed to M.H. (e-mail: holman@cita.utoronto.ca).

Cluster-derived structures and conductance fluctuations in nanowires

R. N. Barnett & Uzi Landman

School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

Understanding the variation of a material's properties with size, form of aggregation and dimensionality is becoming important in the face of increasing miniaturization of electronic and mechanical devices. Experimental studies have focused on the preparation and characterization of solid-state nanometre-scale structures such as metal and semiconductor nanocrystals^{1–3}, surface-supported structures and quantum dots⁴ and nanoscale junctions or wires^{5–22}. It has emerged that these nanostructures can often be fruitfully described using concepts and methodologies developed in the contexts of gas-phase atomic clusters and atomic nuclei^{23–25}. Here we make this connection explicitly through first-principle molecular dynamics simulations^{22,26} which show that, as nanowires of sodium metal are stretched to just a few atoms in diameter, the structures formed by metal atoms in the neck can be described in terms of those observed in small gas-phase sodium clusters²⁷. We find that the electronic spectral and conductance characteristics of these atomic-scale contacts exhibit dynamical thermal fluctuations on a sub-picosecond timescale, owing to rearrangements of the metal atoms, which will significantly affect the transport properties of such nanowires.

Formation and mechanical properties of interfacial junctions (in the form of nanowires) have been predicted through early molecular dynamics simulations⁵, where the materials were modelled using semiempirical embedded-atom potentials. In these studies it was shown that separation of the contact between materials leads to generation of a connective junction which elongates and narrows through a sequence of structural instabilities; at the early stages elongation of the junction involves multiple slip events, whereas at the later stages, when the lateral dimension of the wire narrows to a diameter of ~ 15 Å, further elongation involves a succession of stress accumulation and fast relief stages associated with a

sequence of order–disorder structural transformations localized to the neck region^{5,19,20}. These predictions, as well as anticipated electronic conductance properties^{5,6} have been corroborated in a number of experiments using scanning tunnelling and force microscopy^{5,7–9,11,16,21}, break junctions¹⁰ and pin–plate techniques^{13,19} in ambient environments, as well as under ultra-high vacuum and/or cryogenic conditions.

One of the most fascinating aspects of such nanowires is the prospect of formation of atomic-scale contacts and switches^{5,11,14,18} which may occur towards the ultimate stages of elongation (that is before complete physical separation, or breaking, of the wire). Understanding the atomic rearrangements and structural evolution, electronic properties, and dynamics of such nanojunctions, particularly as they approach the ‘one-atom contact’ regime, is a significant challenge requiring a detailed theoretical description based on first-principles electronic structure calculations. To this end we have used the Born–Oppenheimer (BO) local-spin-density (LSD) molecular dynamics (MD) method (for details of the BO–LSD–MD, see ref. 26) for sodium wires, where the dynamics of the ions (with an integration time-step of 3 fs) evolves at finite temperature ($T = 190$ K) on concurrently calculated ground-state potential-energy surfaces, evaluated self-consistently for each ionic configuration using the LSD theory, in conjunction with non-local norm-conserving pseudopotentials²⁸ with a 6.2 Ry plane-wave energy cut-off.

The simulation starts from a sodium block of bulk body-centred cubic (b.c.c.) (100) layers oriented along the z -axis and tapered to a narrowing in the middle. The periodic super-cell with (x, y, z) dimensions (19.98 Å, 19.98 Å, 21.3 Å) contains a total of 119 atoms treated dynamically, supported at the top and bottom by static (100) layers (25 atoms each); one of these layers is in the cell and the other is its periodic image. After equilibration the structure of the junction is mostly b.c.c., except in the middle layer (Fig. 1a). Each subsequent elongation involves a uniform dilation in the z direction, followed by steepest-descent energy minimization and dynamical equilibration for several picoseconds (at the start of the equilibration period stochastic thermalization to 190 K is used).

The atomic configurations of the sodium wire at selected stages of elongation shown in Fig. 1 reveal development of cluster-derived structures at the narrow-neck region. Particularly striking is the formation of a 13-atom (slightly distorted) icosahedron (5-fold symmetric), supported between the upper and lower parts of the wire (Fig. 1b,c); this structure may also be viewed as two Na₇ pentagonal bipyramidal clusters sharing an apex atom. Subsequent elongation results in ‘opening’ of the structure (Fig. 1d), and stretching of the contact past this stage results in formation of a pentagonal bipyramidal Na₇ cluster (top) bonded through one of its apex atoms to the upper atom of a tetrahedral cluster (bottom), giving the appearance of a stretched sodium dimer at the narrowing (Fig. 1e). Further elongation results in breaking of the nanowire (Fig. 1f). The occurrence of these molecular-like supported structures, which correlate with those derived for isolated sodium clusters²⁷, suggests that in atomic-scale narrow contacts the atomic coordination and nature of bonding may be described using information from investigations pertaining to small (isolated) clusters. In a certain sense the reduced dimensions, increased surface-to-volume ratio, and atomically impoverished environment in such nanoscale solid-state junctions provide a link to the realm of finite-size cluster science, where ground-state (and isomeric) structures are different from the bulk, and ‘magic numbers’ are known to occur^{23–25,27}. (Here ‘magic numbers’ refers to the enhanced stability of certain sequences of cluster sizes, shapes and structural motifs due to electronic and/or geometric-packing shell effects, which may vary from one class of materials to another.)

Contour plots of the self-consistent local effective potential confining the electrons to the wire are shown in Fig. 2 for the 23.97-Å and 29.96-Å wire configurations (compare Fig. 1b and e).

These confinement potential profiles outline the shape of the wire; we note the rather 'smooth' appearance of the 'effective potential tube' (defined by the peripheral outmost solid contour corresponding to the Fermi energy), as a result of effective screening; the approach of the confining potentials to the vacuum level (beyond the outer dashed contour) is similar for both wires, indicating a reduced screening in the more elongated one. In this context we note the close-to-circular shapes of the confining potentials at the narrowing (particularly as the wire narrows; see right-hand panels in Fig. 2), which correlates with the conductance quantization pattern⁶ reported for sodium wires¹⁰.

We now examine the electronic spectral properties of these nanowires, and their transport characteristics. The time-averaged global and local (in the narrowing) electronic densities of states (DOS), obtained from a dynamical simulation at 190 K of the

nanowire corresponding to the selected configuration shown in Fig. 1b, is displayed in Fig. 3a, along with the time-variation of the DOS at the Fermi level, $N(\epsilon_F)$ in Fig. 3b). We observe that while both the global and local $N(\epsilon_F)$ are finite, indicating metallic character, they exhibit significant fluctuations on a subpicosecond timescale (Fig. 3b), originating from redistribution of the electronic energy levels caused by ionic motion. Such fluctuations are expected to be portrayed in the electronic transport through the nanowire.

Using the density-functional Kohn–Sham orbitals the electronic conductivity, $\sigma(\omega)$, can be calculated from the linear-response (Kubo) formula^{29,30}

$$\sigma(\omega) = \frac{2\pi e^2 \hbar}{\Omega m} \sum_{ij} \frac{(f_i - f_j)}{\epsilon_i - \epsilon_j} \left| \langle i | \hat{p}_z | j \rangle \right|^2 \delta(\epsilon_i - \epsilon_j - \hbar\omega) \quad (1)$$

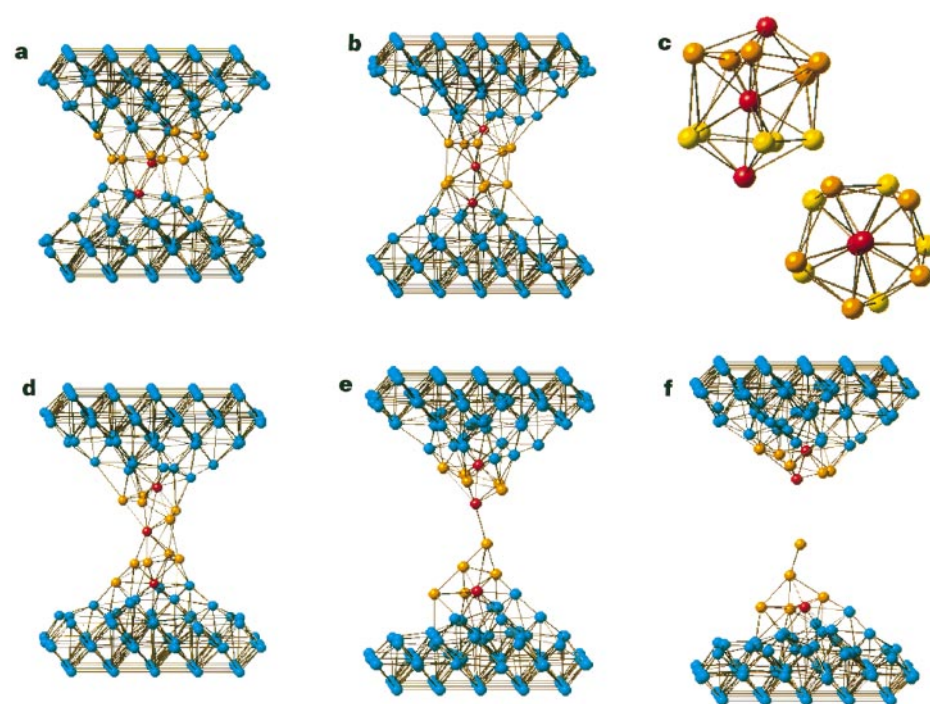


Figure 1 Atomic configurations of a sodium nanowire at selected elongation stages, obtained through first-principles molecular dynamics simulations. The configurations in **a**, **b**, **d**, **e** and **f** correspond to wires of lengths 21.3 Å, 23.97 Å, 26.64 Å, 29.96 Å and 31.96 Å, respectively. In **c**, side and top views are shown of the stretched icosahedral supported cluster (see neck region in **b**). The 13 atoms forming the icosahedron are distinguished also by colour in the other configurations to allow visualization of the structural evolution.

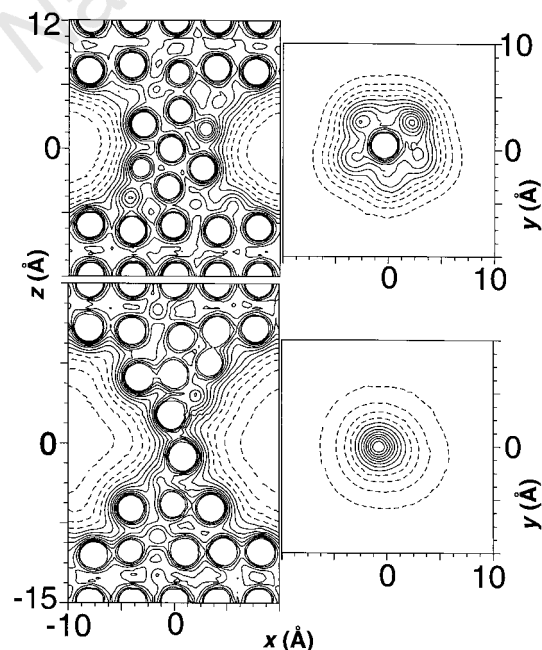


Figure 2 Contour plots of the local part of the self-consistent LSD effective potentials corresponding to the wire configurations shown in Fig. 1b (top panel) and Fig. 1e (bottom panel); contours inside the repulsive atomic cores, where the non-local contribution is significant, are not shown. On the left, contours are in the plane containing the axis (z) of the wire; on the right they are drawn in the transverse plane at the narrowest part of the wire. All contours, with a spacing of 0.68 eV, correspond to energies below the vacuum level, with the solid lines corresponding to energies below ϵ_F ; from the difference between the vacuum level and ϵ_F the work function of the wires is estimated as ~ 2.1 eV.

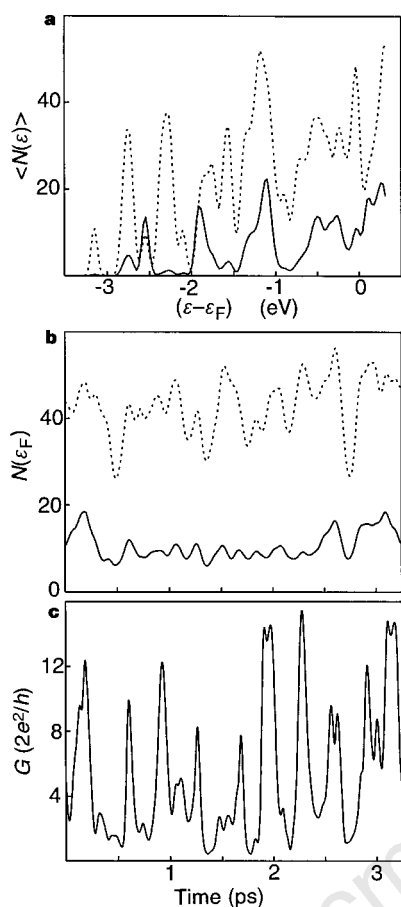


Figure 3 Spectral and conductance characteristics for the 23.97-Å nanowire (Fig. 1b). **a**, Time averages (from simulations at 190 K) of the global (dashed line) and local (solid line; evaluated at the narrowest, cluster, region) densities of states, $\langle N(\epsilon) \rangle$. **b**, Time variations of the global and local DOS at ϵ_F . **c**, Time variation of the conductance G , showing sub-picosecond fluctuations, with an average value of $(5.2 \pm 0.7)g_0$. The Kubo conductivity formula (equation (1), main text) used in our calculations has been discussed previously^{29,30} and its use in conjunction with finite-temperature simulations alleviates to a large extent some of the issues discussed there. In our calculations the delta-function is represented by a gaussian with a width of $\eta = 0.0163$ eV ($\eta/k_B = 190$ K), which is equal to half the average level spacings near ϵ_F . Larger broadening decreases, but does not obliterate, the magnitude of the conductance fluctuations.

where m is the electron mass, f_i is the occupation (Fermi function) for the LSD Kohn–Sham level of energy ϵ_i , $\hat{p}_z = -i\hbar\partial/\partial z$, and Ω is the volume of the periodic calculational cell ($\Omega = AL$, where A and L are, respectively, the periodic-cell transverse area, and length along the nanowire axis). Extrapolation of the time-averaged conductance $\langle G(\omega) \rangle = \langle \sigma(\omega) \rangle A/L$ to $\omega \rightarrow 0$, yields for the 23.96-Å nanowire a d.c. conductance $\langle G \rangle = (5.2 \pm 0.7)g_0$, where $g_0 = 2e^2/h$ ($g_0^{-1} = 12.9$ k Ω) and the statistical error is estimated following ref. 31. Similar analyses yield conductance values of $(11.0 \pm 1.5)g_0$ and $(1.2 \pm 0.3)g_0$ respectively for the 21.3-Å and 26.64-Å wires (Fig. 1a and d); for the 29.96-Å wire (Fig. 1e), a value much smaller than $1g_0$ is predicted (a similar structure and conductance were obtained also for a 28.1-Å wire). The conductance shows an expected decreasing trend as the wire narrows; we note here that no attempt has been made here to search for wire configurations (that is elongations) that may yield a particular sequence of conductance values. These results, as well as the lateral shapes of the effective confinement at the narrowing (Fig. 2) may be compared with phenomenological quantum and semi-classical theories of electronic conductance in wires.^{6,15,17–20,32} In such theories, the confinement of the nano-

constriction is often modelled using hard-wall potentials (or variations thereof, including finite height of the confining potential and electron density spill-out effects), and the conductance channels (in the framework of a Landauer theory) are determined as the lateral electronic states (including their degeneracies) calculated at the narrowest part of the constriction. Our calculated average conductance values are in good correspondence with such semi-classical estimates (except for the lowest one where such estimates are not reliable¹⁵). For example, using the Weyl-modified Sharvin formula corrected for the finite height of the confining potential (equation (3) in ref. 15, with $\alpha = 0.1$ and $k_F = 0.92$ Å⁻¹) gives a value of $5.2g_0$ for a cylindrical wire with a cross-sectional radius of 5.2 Å. This compares favourably with an effective radius of 4.9 Å determined from the confining potential contour (at ϵ_F) corresponding to the 23.97-Å wire (see upper-right cross-sectional view in Fig. 2) for which, as described above, we obtain through equation (1) $\langle G \rangle = (5.2 \pm 0.7)g_0$.

In light of our above-mentioned observation of temporal fluctuations in $N(\epsilon_F)$, caused by finite-temperature dynamical structural fluctuations, we show in Fig. 3c the corresponding calculated time variations of G (whose time-average yields the value of $5.2g_0$). The resonance-like features in G correlate with the temporal structural and electronic spectral fluctuations (in both the local density of states and momentum matrix elements) due to thermal vibrations. Such finite-temperature effects and factors related to the ‘non-ideal’ (structural) nature of the nanowires underlying scattering of the electrons, as well as inherent irreproducibility in the fabrication of the nanowires through contact elongation or break junction techniques, are likely to be the sources of the distribution of measured conductance values (commonly displayed as conductance histograms of G). Consequently, our results suggest that dynamical conductance fluctuations may be of importance in considerations pertaining to atomic-scale contacts, and that they might be identified experimentally by fast spectroscopical measurements of electronic transport in nanowires. □

Received 13 February; accepted 6 May 1997.

- Whetten, R. L. *et al.* Nanocrystal gold molecules. *Adv. Mater.* **8**, 428–433 (1996).
- Luedtke, W. D. & Landman, U. Structure, dynamics and thermodynamics of passivated gold nanocrystallites and their assemblies. *J. Phys. Chem.* **100**, 13323–13329 (1996).
- Alivisatos, A. P. Semiconductor clusters, nanocrystals and quantum dots. *Science* **271**, 933–937 (1996).
- Avouris, P. (ed.) *Atomic and Nanometer-Scale Modification of Materials: Fundamentals and Applications* (Kluwer, Dordrecht, 1993).
- Landman, U., Luedtke, W. D., Burnham, N. & Colton, R. J. Atomistic mechanisms and dynamics of adhesion, nanoindentation and fracture. *Science* **248**, 454–461 (1990).
- Bogachev, E. N., Zagorin, A. M. & Kulik, I. O. Conductance jumps and magnetic flux quantization in ballistic point contacts. *Sov. J. Low Temp. Phys.* **16**, 796–800 (1990); *Fiz. Nizk. Temp.* **16**, 1404–1411 (1990).
- Pascual, J. I. *et al.* Quantum contact in gold nanostructures by scanning-tunneling microscopy. *Phys. Rev. Lett.* **71**, 1852–1855 (1993).
- Olesen, L. *et al.* Quantized conductance in an atom-sized point contact. *Phys. Rev. Lett.* **72**, 2251–2254 (1994).
- Pascual, J. I. *et al.* Properties of metallic nanowires: from conductance quantization to localization. *Science* **267**, 1793–1795 (1995).
- Krans, J. M., van Ruitenbeek, J. M., Fisun, V. V., Yanson, I. K. & de Jongh, L. J. The signature of conductance quantization in metallic point contacts. *Nature* **375**, 767–769 (1995).
- Smith, D. P. E. Quantum point contact switches. *Science* **269**, 371–373 (1995).
- Bratkovsky, A. M., Sutton, A. P. & Todorov, T. N. Conditions for conductance quantization in realistic models of atomic-scale metallic contacts. *Phys. Rev. B* **52**, 5036–5051 (1995).
- Costra-Kramer, J. L., Garcia, N., Garcia-Mochales, P. & Serena, P. A. Nanowire formation in macroscopic metallic contacts: quantum mechanical conductance tapping a table top. *Surface Sci.* **342**, 11144–11449 (1995).
- Lang, N. D. Resistance of atomic wires. *Phys. Rev. B* **52**, 5335–5342 (1995).
- Garcia-Martin, A., Torres, J. A. & Saenz, J. J. Finite size corrections to the conductance of ballistic wires. *Phys. Rev. B* **54**, 13448–13451 (1996).
- Rubio, G., Agrait, N. & Vieira, S. Atomic-sized metallic contacts; mechanical properties and electronic transport. *Phys. Rev. Lett.* **76**, 2302–2305 (1996).
- Scherbakov, A. G., Bogachev, E. N. & Landman, U. Quantum electronic transport through three-dimensional microconstrictions with variable shapes. *Phys. Rev. B* **53**, 4054–4064 (1996).
- Bogachev, E. N., Scherbakov, A. G. & Landman, U. Magnetic switching and thermal enhancement of quantum transport through nanowires. *Phys. Rev. B* **53**, R13246–R13249 (1996).
- Landman, U., Luedtke, W. D., Salisbury, B. E. & Whetten, R. L. Reversible manipulations of room temperature mechanical and quantum transport properties in nanowire junctions. *Phys. Rev. Lett.* **77**, 1362–1365 (1996).
- Landman, U., Luedtke, W. D. & Gao, J. Atomic-scale issues in tribology: interfacial junctions and nano-elastohydrodynamics. *Langmuir* **12**, 4514–4528 (1996).
- Stalder, A. & Durig, U. Study of yielding mechanics in nanometer-sized Au contacts. *Appl. Phys. Lett.* **68**, 637–639 (1996).

22. Landman, U., Barnett, R. N. & Luedtke, W. D. Nanowires: size evolution reversibility and one-atom contacts. *Z. Physik D* (in the press).
23. de Heer, W. A. The physics of simple metal clusters: experimental aspects and simple models. *Rev. Mod. Phys.* **65**, 611–675 (1993).
24. Martin, T. P. (ed.) *Large Clusters of Atoms and Molecules* (Kluwer, Dordrecht, 1996).
25. Yannouleas, C. & Landman, U. in *Large Clusters of Atoms and Molecules* (ed. Martin, T. P.) 131–200 (Kluwer, Dordrecht, 1996).
26. Barnett, R. N. & Landman, U. Born-Oppenheimer molecular-dynamics simulations of finite systems: structure and dynamics of $(\text{H}_2\text{O})_2$. *Phys. Rev. B* **48**, 2081–2097 (1993).
27. Haberland, H. (ed.) *Clusters of Atoms and Molecules* (Springer Ser. in Chem. Phys. 52 & 57, Springer, Berlin, 1994).
28. Troulier, N. & Martin, J. L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **43**, 1993–2006 (1991).
29. Thouless, D. J. & Kirkpatrick, S. Conductivity of the disordered linear chain. *J. Phys. C* **14**, 235–245 (1981).
30. Imry, Y. & Shiren, N. Energy averaging and the flux-periodic phenomena in small normal-metal rings. *Phys. Rev. B* **33**, 7992–7997 (1986).
31. Flyvbjerg, H. & Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **91**, 461–466 (1989).
32. Bogachev, E. N., Scherbakov, A. G. & Landman, U. Shape-effects on conductance quantization in three-dimensional nanowires: hard versus soft potentials. *Phys. Rev. B* (in the press).

Acknowledgements. Calculations were performed at the National Energy Research Scientific Computing Center, Lawrence Berkeley, CA, and the GIT Center for Computational Materials Science. This work was supported by the DOE and AFOSR.

Correspondence should be addressed to U.L. (e-mail: uzi.landman@physics.gatech.edu).

Hollow nanoparticles of WS_2 as potential solid-state lubricants

L. Rapoport*, Yu. Bilik*, Y. Feldman†, M. Homyonfer†, S. R. Cohen‡ & R. Tenne†

* Department of Mechanics and Control, Center for Technological Education-Holon, PO Box 305, Holon 58102, Israel

† Department of Materials and Interfaces, ‡ Chemical Services Unit, Weizmann Institute, Rehovot 76100, Israel

Solid lubricants fill a special niche in reducing wear in situations where the use of liquid lubricants is either impractical or inadequate, such as in vacuum, space technology or automotive transport. Metal dichalcogenides MX_2 (where M is, for instance, Mo or W and X is S or Se) are widely used as solid lubricants. These materials are characterized by a layered structure with weak (van der Waals) inter-layer forces that allow easy, low-strength shearing^{1,2}. Within the past few years, hollow nanoparticles (HNs) of MX_2 with structures similar to those of nested carbon fullerenes and nanotubes have been synthesized^{3,4}. Here we show that these materials can act as effective solid lubricants: HN- WS_2 outperforms the solid lubricants 2H- MoS_2 and 2H- WS_2 in every respect (friction, wear and lifetime of the lubricant) under varied test conditions. We attribute the outstanding performance of HN- WS_2 to its chemical inertness and the hollow cage structure, which imparts elasticity and allows the particles to roll rather than to slide.

The material used in this study was synthesized by a solid-gas reaction⁵ which yields gram quantities of HN- WS_2 , with limited control over the size distribution and shape of the nanoparticles. The average particle size of the 2H- WS_2 and 2H- MoS_2 powders was 4 μm ; the average particle size of HN- WS_2 was 120 nm. Samples of the 2H- MoS_2 and 2H- WS_2 powders were milled for 24 hours, leading to platelets with an average size of 0.5 μm . Figure 1 shows schematic illustrations of the two main experimental set-ups which were used in the present work to evaluate the tribological properties of HN- WS_2 in comparison to bulk 2H- WS_2 (and 2H- MoS_2) powder. Using the disk-block tester at high sliding speeds and low loads (experiment A in Table 1), the rolling friction efficacy of HN- WS_2 can be examined. At high loads and lower velocities, the disk-block tester (experiments B and C in Table 1) probes the compliance of the HN films under conditions where mechanical effects dominate.

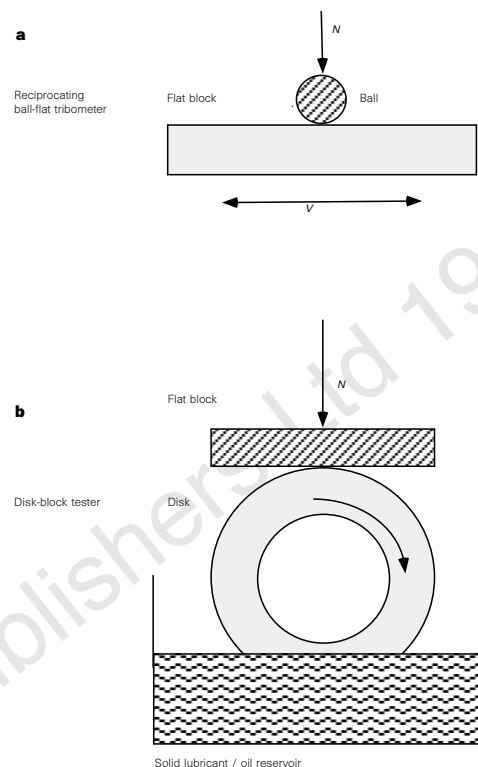


Figure 1 Diagrams of the two experimental set-ups which were used in the present work to investigate the tribological properties of HN- WS_2 and 2H- WS_2 (2H- MoS_2). **a**, Ball-flat tribometer: here, friction and wear of the various lubricants were investigated under low sliding velocities and loads. The load (N) and velocity (V) of the flat were varied in a predetermined manner and the friction coefficient (slope of friction versus load) was determined. The size of the wear track and the ball surface were examined after each run with an optical microscope. Quantitative analysis of the wear tracks with scanning force microscope (SFM) was undertaken. To test the solid lubricants under more realistic conditions (larger loads and velocities), the disk-block tester **b** was employed.

Figure 2 shows the results of the wear measurements using the tribometer. Wear volumes were determined from the depth and width of the wear track. Wear volumes of 2H- MoS_2 both dry and wetted with a drop of oil, were 6.5 and 2 times that of HN- WS_2 , respectively. The wear for pure mineral oil is 65 times greater. These results (supported by other data; see Supplementary Information) confirm the visual impression of Fig. 2, namely that the wear track produced by the wall when HN- WS_2 powder was used, was the smallest of all solid lubricants examined. Although the ball surface was under continuous load during the entire test, each segment of the flat block experienced only instantaneous load as the ball slid over it. Therefore the ball surface suffered most of the thermal stress, which is shown by the large difference in its wear among the different powders (Fig. 2). Here again, the amount of wear for the HN- WS_2 powder was smallest.

In Table 1 we show the results of an experiment in which rolling of the particles is allowed, because of high speeds and low loads (experiment A). To simulate typical industrial conditions, only 5 wt% solid lubricant, in a commercial oil, was used. We note that friction (μ) and wear (w) coefficients, and also the average surface roughness (R_a), were smallest for the HN (see also Supplementary Information). In experiments B and especially, experiment C (Table 1) the compressive regime is accessed. Obviously, the wear of the HN- WS_2 /oil mixture was the smallest of the three, which demonstrates again the superior wear protection of the nano-material compared with the 2H counterparts. It is also clear that

Table 1 Results of wear experiments

Exp.	Velocity (m s^{-1})	Load (N)	Conc. (wt%)	Coeff.	Pure oil	2H-MoS ₂ (4 μm)	2H-WS ₂ (0.5 μm)	2H-WS ₂ (4 μm)	HN-WS ₂ (120 μm)
A	0.44	300	5	μ	0.07	0.07		0.05	0.03
				w	1.6×10^{-8}	1.3×10^{-8}		1.5×10^{-8}	0.7×10^{-8}
B	0.22	1,800	20	R_a	0.83	0.75		1.18	0.53
				w		7.9×10^{-11}		2.3×10^{-11}	5.1×10^{-11}
				w		3.9×10^{-11}		0.043	1.3×10^{-11}
				μ			0.10	0.042	0.034
C	0.11–0.44	300–3,000	60	μ			0.067		0.028
				$T^\circ\text{C}$			4.6×10^{-4} 88	2.8×10^{-4} 72	1.9×10^{-4} 72

Wear (w) and friction (μ) coefficients of a steel block in contact with a steel disk for four kinds of solid lubricants mixed with mineral oil. Average particle size shown in parentheses at the head of each column. In experiment A, the sliding track length was 1.27×10^4 m (8 h). A commercial oil for transmission systems (Delcol) was used. w is given in units of $\text{mm}^3 \text{mm}^{-1} \text{N}^{-1}$. Profiles of the wear track were measured by stylus profilometry. The results of the experiment with the powder of 2H-MoS₂ with 0.5 μm grain size was very similar to the results of the 4- μm 2H-MoS₂ powder and is therefore not reported in detail. The average roughness (R_a ; μm) of the area of the wear-track was determined after the experiment and is also included. In experiments B, the sliding track length was 2.38×10^3 m (6 h). In experiment C, the load on the disk was increased from 300 N to 3,000 N in steps of 300 N, each step lasting 300 s. At each load, the velocity was increased from 0.11 to 0.44 m s^{-1} in steps of 0.11 m s^{-1} (3.5 h per experiment). The temperature of the flat block was determined during the experiment by contacting a thermocouple to the block. Reported temperature is average after the initial run-in.

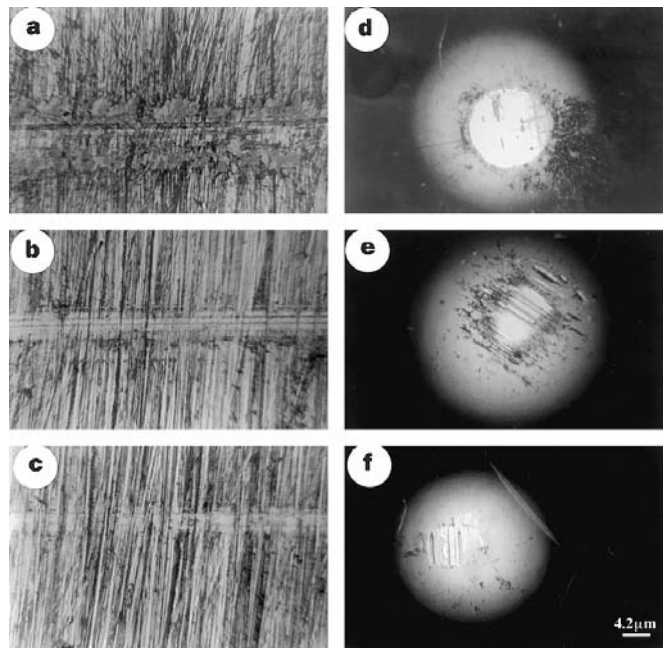


Figure 2 A series of images from the ball-flat tribometer experiment demonstrating the influence of various solid lubricants on friction and wear of the flat block and the ball: 2H-WS₂ (**a** and **d**); 2H-MoS₂ (**b** and **e**); HN-WS₂ (**c** and **f**). Load, 150 g; number of cycles, 150; velocity of the flat, $2 \times 10^{-4} \text{ m s}^{-1}$ (wear track 1 m, duration 4 h). The flat surface was made of tool steel AISI 01, and the ball was made of quenched and tempered steel–AISI 52100 (Vickers hardness 220 and 600, respectively). To avoid run-away of the powder (especially the HNs) during the experiment, the flat metal surface was abraded with glass paper and one drop of mineral oil was added to 80 mg of solid lubricant. When the experiment was run for 500 cycles (12 h) the wear for all samples increased, while the difference in wear between samples remained the same.

w decreases with increasing fraction of solid lubricant in the mixture. Experiment C represents an extended experiment which tests wear under varying speeds and loads (ASTM no. G77)⁶. The results show reduction in wear for the HNs, relative to both sizes of 2H-WS₂ platelets.

We now summarize another set of experiments that are fully described in the Supplementary Information. Single-crystal LiF is routinely used for determination of the dislocation density induced by a top sliding surface. In the present experiment, LiF crystals were immersed in mixtures containing 5% of the solid lubricants for 5 min and then used as the flat surface in the tribometer (Fig. 1a). Whereas little damage was inflicted on the LiF surface in the case of HN-WS₂, severe ‘ploughing’ of the surface was observed when 2H-MoS₂ was used. This experiment directly measures shear stress, thus providing further support for the superiority of the HN lubricant layer.

The desirable tribological properties of MoS₂ films can be attributed to the weak van der Waals forces operating between the layers, which leads to easy shear of the films with respect to each other^{7,8}. This shear mechanism is not likely to control the tribological properties of the HN-WS₂ powder, as appreciable shear between the inner and outer closed layers is not possible. The tribological benefits of the HNs probably arise from their unique

properties, as outlined here. The particle size of the HNs, typically >30 nm, is sufficient to prevent asperity contact between the mating surfaces. This, taken together with their round shape, opens the possibility of rolling friction. Further, elastic as opposed to inelastic deformations diminish the energy dissipation associated with friction and wear under the load. The hollow cage structure of the HNs imparts a high elasticity which augments their resilience⁹.

Chemical effects can also play an important role in wear protection^{1,7}. The endurance of sputtered MoS₂ films under wear conditions has been found to be better in dry argon gas⁸ or in vacuum¹⁰ than in air, indicating a chemical reaction in which MoS₂ oxidizes to MoO₃ or more complex mixed metal oxides. The chemical reactions that are relevant to wear of MoS₂ platelets occur predominantly at the prismatic edges (llc), where reactive dangling bonds exist. The absence of dangling bonds may therefore be one of the prime advantages of HNs over the crystalline platelet (2H) particles for reduction of friction and wear. A further, related benefit is lowered adhesion to the underlying substrate.

Microscopic measurements were performed to substantiate the hypothesis that adhesion at the dangling bonds mediates the friction in metal dichalcogenides (see Supplementary Information). Using friction force microscopy¹¹ and a modified Hertzian treatment to estimate contact area¹², we found that the step edges on

single-crystal MoS₂ have over four times the shear strength of the terraces. The weak adhesion of the HNs to the mating surface were reflected in scanning force microscopy (SFM), whereby the HN particles were brushed away by the tip in contact mode imaging. Using the non-contact mode of the SFM, which avoids shear, clear images of the HN were obtained. Transmission electron microscope (TEM) analysis of the HNs which had gone through extensive wear experiments revealed that many of them had lost their outer WS₂ shells, but that they had nevertheless preserved their overall cage shape. In contrast, the 2H material underwent extensive structural damage, amorphization and oxidation. Consequently, improved tribological behaviour of the HNs, attributed to rolling friction and absence of dangling bonds, is not lost until complete destruction of the particles occurs. The slower deterioration of the HNs is reflected by their lower wear rates, even after 12 hours of sliding (Fig. 2 legend).

The experimental results (rows 6, 7 and 8 of Table 1) clearly show that both μ and w increase with decreasing size of the 2H platelets. This observation is in agreement with the suggestion that dangling bonds mediate friction in metal dichalcogenides: as the average size of the platelets decreases, the total surface area and hence the amount of dangling bonds on the reactive (11c) surface increases. In contrast, the friction coefficient of the much smaller HN particles, which do not expose dangling bonds, is less than that of any of the different 2H particle mixtures. A decrease in particle size also tends to increase viscosity, which usually entails higher friction¹³. The HN mixture had the highest viscosity, and normalizing for this parameter would therefore give an even stronger indication of the enhanced performance of the HN material, relative to 2H-MoS₂ and 2H-WS₂.

The synthesis of HN-WS₂ uses abundant precursors and fairly straightforward conditions, and the cost of the material is therefore not expected to prohibit large-scale applications in the automotive and aerospace industry, which currently use the apparently inferior dichalcogenides. Improved control of the size distribution and shape of the nanoparticles is expected to lead to improved tribological properties. Furthermore, HN-MoS₂ might be expected to outperform HN-WS₂, by analogy with their platelet counterparts. Clearly, alternative synthetic routes for the production of HN materials could yield solid lubricants with further improved properties. □

Received 6 December 1996; accepted 21 April 1997.

1. Singer, I. L. in *Fundamentals of Friction: Macroscopic and Microscopic Processes* (eds Singer, I. L. & Pollock, H. M.) 237 (Kluwer, Dordrecht, 1992).
2. Bowden, F. P. & Tabor, D. *Friction: An Introduction to Tribology* 91 (Anchor, Garden City, New York, 1973).
3. Tenne, R., Margulis, L., Genut, M. & Hodes, G. Polyhedral and cylindrical structures of tungsten disulphide. *Nature* **360**, 444–445 (1992).
4. Feldman, Y., Wasserman, E., Srolovitz, D. J. & Tenne, R. High-rate, gas-phase growth of MoS₂ nested inorganic fullerenes and nanotubes. *Science* **267**, 222–225 (1995).
5. Feldman, Y. *et al.* Bulk synthesis of inorganic fullerene-like MS₂ (M=Mo, W) from the respective trioxide and the reaction mechanism. *J. Am. Chem. Soc.* **118**, 5362–5367 (1996).
6. *Standard G77 in ASTM Book of Standards Vol. 3.02, Wear and Erosion: Metal Corrosion* 303–314 (Am. Soc. for Testing and Materials, West Conshohocken, PA, 1993).
7. Seitzman, L. E., Bolster, R. N., Singer, I. L. & Wegand, J. C. Relationship of endurance to microstructure of MoS₂ coatings. *Tribology Trans.* **38**, 445–451 (1995).
8. Moser, J. & Lévy, F. MoS₂ lubricating films: structure and wear mechanisms investigated by cross-sectional transmission electron microscopy. *Thin Solid Films* **228**, 257–260 (1993).
9. Srolovitz, D. J., Safran, S. A., Homyonfer, M. & Tenne, R. Morphology of nested fullerenes. *Phys. Rev. Lett.* **74**, 1779–1781 (1995).
10. Fischer, T. E. in *Fundamentals of Friction: Macroscopic and Microscopic Processes* (eds Singer, I. L. & Pollock, H. M.) 299 (Kluwer, Dordrecht, 1992).
11. Meyer, E. *et al.* Site-specific friction force spectroscopy. *J. Vac. Sci. Technol. B* **14**, 1285–1288 (1996).
12. Mamin, H. J., Ganz, E., Abraham, D. W., Thomson, R. E. & Clarke, J. Contamination-mediated deformation of graphite by the scanning tunneling microscope. *Phys. Rev. B* **34**, 9015–9018 (1986).
13. Klein, J. Shear, friction, and lubrication forces between polymer-bearing surfaces. *Annu. Rev. Mater. Sci.* **26**, 581–612 (1996).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank D. J. Srolovitz for discussions. This work was supported in part by the ACS-PRF; the UK-Israel Science and Technology Research Fund; the Minerva Foundation (Munich); and the Israel Ministry of Science (Strategic Program on Nanomaterials).

Correspondence and requests for materials should be addressed to R.T. (e-mail: cpreshef@weizmann.ac.il).

Low nitrate : phosphate ratios in the global ocean

T. Tyrrell* & C. S. Law†

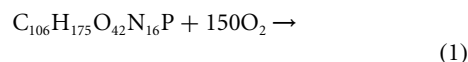
* Department of Oceanography, Southampton Oceanography Centre, University of Southampton, European Way, Southampton SO14 3ZH, UK

† Plymouth Marine Laboratory, Prospect Place, Plymouth PL1 3DH, UK

The distribution and availability of the nutrients nitrate and phosphate exert a strong control over primary production in the world ocean. Here we use a recently compiled global data set of oceanographic observations¹—a database that is at least 40 times larger than currently used data sets (for example, ref. 2)—to analyse the variation of the nitrate : phosphate concentration ratio with depth and geographical location. Although the nutrient distributions confirm the dominant influence of aerobic decomposition, in agreement with the observations of Redfield³, we also identify a hitherto unreported secondary trend at low nitrate : phosphate ratio (~ 2 –3). These conditions of low nitrate : phosphate ratio are associated with low oxygen concentrations and are probably caused by denitrification. Examination of the geographical distribution of these low nitrate : phosphate data suggests that denitrification in the western and northern North Pacific Ocean may have been previously overlooked, and that a reassessment of the global oceanic denitrification budget may therefore be required.

The World Ocean Atlas 1994 (WOA94) data set¹ is an international collection of oceanographic information covering the past 60 years. Part of this data set has previously been used to construct global contour plots of nitrate, phosphate and silicate concentrations at several depths⁴. Here the 681,381 points in the data set are used to produce a scatter plot (Fig. 1) of the concentrations of nitrate (NO₃⁻) versus phosphate (PO₄³⁻), which is then analysed. Despite the broad spatial and temporal range of the data, three features are clearly apparent. (1) A strong trend along a line of slope $[\text{NO}_3^-] \approx (14\text{--}15)[\text{PO}_4^{3-}]$. (2) A positive x -axis intercept of this main trend, that is, a predominance of points near to the $[\text{NO}_3^-] = 0$ axis rather than near to the $[\text{PO}_4^{3-}] = 0$ axis. (3) A weaker trend along lines with slopes $[\text{NO}_3^-] \approx 2[\text{PO}_4^{3-}]$. A least absolute deviation best-fit line⁵ to all of the data except the weaker trend (defined below in equation (2)) actually produces $[\text{NO}_3^-] = 14.1[\text{PO}_4^{3-}] - 1.53$ (with a mean absolute $[\text{NO}_3^-]$ deviation of each point from the best-fit line of $3.15 \mu\text{mol NO}_3^- \text{ kg}^{-1}$), although the data is biased towards surface waters (where there has been more sampling). Analysis of only the points below 1,000 m (again excluding points in the weaker trend) reveals an average deep-sea $[\text{NO}_3^-] : [\text{PO}_4^{3-}]$ ratio of 14.6.

The main trend in the data with slope $[\text{NO}_3^-] \approx (14\text{--}15)[\text{PO}_4^{3-}]$ is due to phytoplankton assimilation of the nutrients in surface waters followed by aerobic decomposition of the sinking organic matter³, which recycles nitrogen and phosphorus back into dissolved form in deeper water according to the reaction⁶:



Some scatter about a straight line is to be expected due to variation in the nitrate : phosphate ratios of the plankton being decomposed⁶, localized riverine input and other factors. The positive x -intercept of the main trend in the data agrees with the observed predominance of nitrogen over phosphorus limitation at low nutrient concentrations in surface waters^{7,8}.

The most surprising aspect of the data is the weaker trend along lines with slopes $[\text{NO}_3^-] : [\text{PO}_4^{3-}] \approx 2.0$. This trend can be seen

clearly in a profile of $[\text{NO}_3^-] : [\text{PO}_4^{3-}]$ against depth (Fig. 2). This feature is not remarked upon in previous nitrate:phosphate analyses^{3,9–11} because it is not apparent in the smaller data sets² which these studies have used.

To investigate this new phenomenon, a subset of the data containing the points contributing to the secondary nitrate:phosphate trend was extracted according to the following criteria:

$$[\text{PO}_4^{3-}] > 1.5 \mu\text{mol kg}^{-1} \text{ and } [\text{NO}_3^-]/[\text{PO}_4^{3-}] < 3.0 \quad (2)$$

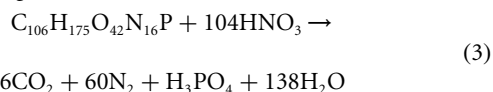
(the shaded area in Fig. 1 inset). These criteria were chosen so as to include points at low nitrate:phosphate, but to exclude points that are only a short 'distance' away from the predominant trend in the data. For example, only a minor system perturbation or measurement error is required at lower nutrient concentrations for the nitrate:phosphate ratio to change from 15 to 2, and so our criteria only include low nitrate:phosphate ratios at higher phosphate concentrations. The low nitrate:phosphate (LNP) points resulting from this definition make up 1.8% of the total (12,469 points out of 681,381). The geographical distribution of these LNP points can be plotted (Fig. 3). Relationships with other parameters (Fig. 4) show whether the LNP points occur preferentially under certain conditions or at certain times.

The first explanation to consider for the LNP points is whether they are due solely to instrument error, for example, associated with older nutrient measurement techniques. However, there is no strong pattern in the yearly distribution of the LNP points (Fig. 4a), with a large fraction (37% = 4,599) of them being made during the past 20 years. In addition, the LNP measurements were made during cruises by many different nations, using different methodologies, in almost all oceans, and across a range of depths. LNP

points are detected by cruises visiting the same area many years apart. For instance, LNP points have been detected independently in the western North Pacific (110–180°E, 0–60°N) by Japanese, Taiwanese, USA, Chinese, USSR, Thai and Indonesian cruises in 1949, 1958 and in 21 out of the 24 years between 1963 and 1986. LNP points and normal (nitrate:phosphate ≈ 15) points were often measured at different depths within the same profile. There is strong spatial coherence (clustering) of the LNP points (Fig. 3), and they possess a spatial distribution which is different from that of the data set as a whole. There is a definite correlation with oxygen concentrations (Fig. 4b). Instrument error can be discounted as a possible cause of the LNP points.

Because the low nitrate:phosphate ratios are genuine, what could have caused them? Examination of global nitrate:silicate and silicate:phosphate plots (not shown) reveal a low trend in the former but not a low trend in the latter, so the low nitrate:phosphate ratios are caused by low nitrate, not high phosphate. As the nitrate:phosphate ratio of river water is generally greater than that of the Redfield ratio¹², and values of this ratio in phytoplankton biomass of <3 have not been reported⁶, this suggests that a nitrate removal process is responsible for the LNP points.

In particular, denitrification is usually assumed to be the explanation for low nitrate:phosphate ratios^{9,11,13}. Anaerobic decomposition by denitrifying bacteria alters the nutrient concentrations in the water according to a reaction similar to^{14,15}:



Denitrification is more likely to dominate in low-oxygen waters, although it can still be operative in more oxic waters within the

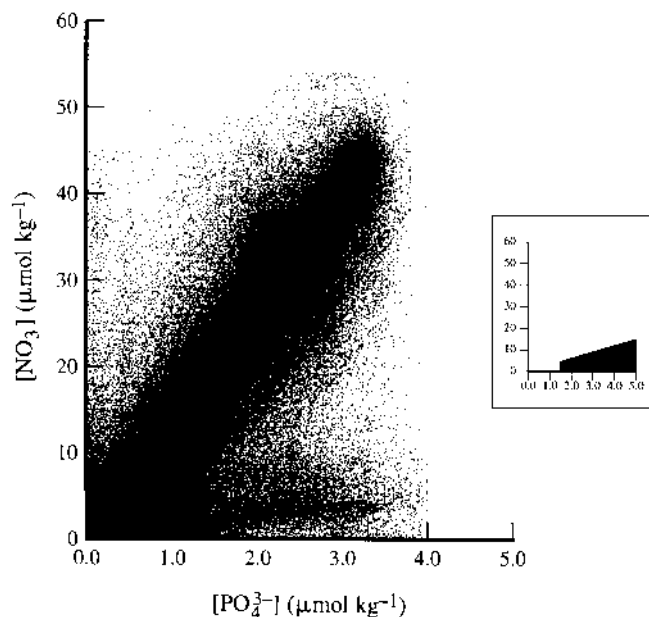


Figure 1 Scatter plot of nitrate concentration versus phosphate concentration in the global ocean, with the data points taken from ref. 1. Data were subjected to screening procedures by the National Oceanographic Data Center (NODC) before being included; historical data sets were recovered and included subject to reliability³⁰. The data comes from 68,341 different stations during cruises by 51 different countries. Only data points with no error flags are used here. Only actual measured values are used, with no interpolations being included. Observations were only included when both nitrate and phosphate were measured simultaneously, that is from the same bottle. The upper limit for phosphate in ref. 1 is $4.0 \mu\text{mol kg}^{-1}$. The shaded area in the inset shows the subset of the data designated as low nitrate:phosphate ratio (LNP ratio).

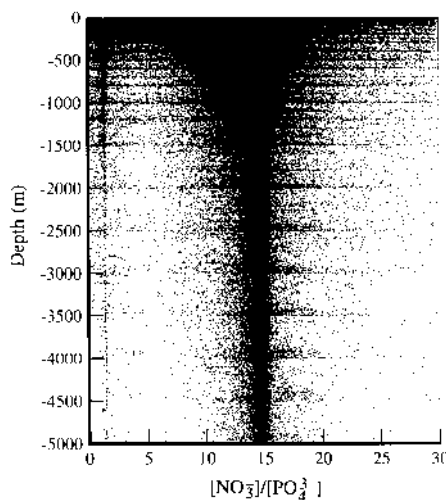


Figure 2 Scatter plot of the nitrate:phosphate ratio versus depth in the global ocean (after ref. 10), using the same data as in Fig. 1.

anoxic micro-sites in particulate matter¹⁶. The geographical distribution of LNP points (Fig. 3) provides further supporting evidence of denitrification as the source process, with LNP points predominantly observed in low-oxygen-prone coastal areas and regions of open-ocean low oxygen associated with upwelling waters.

The LNP points correspond to large nitrate deficits of between 18 and 45 $\mu\text{mol kg}^{-1}$ relative to nitrate:phosphate = 15, and the nitrogen in the missing nitrate must have been converted into some other form of nitrogen. Neither ammonium or nitrite¹⁷ are detected at such high concentrations in oxic or sub-oxic waters. However, N_2 , the end-product of denitrification, is present at concentrations of $\sim 1,200 \mu\text{mol kg}^{-1}$ due to equilibration with

the large atmospheric reservoir. Large amounts of nutrients may also exist in dissolved organic form (dissolved organic nitrogen, DON, and dissolved organic phosphorus, DOP), but there is no reason to suspect an increase in DON without a simultaneous increase in DOP, or that this should be correlated with low oxygen. Other nitrogen species in the oceans are only present at trace concentrations, and so conservation of nitrogen also suggests that a flux of NO_3^- to N_2 via denitrification is the most likely cause of the shortfalls in NO_3^- in the LNP points.

From the arguments above, removal of NO_3^- by denitrification seems by far the most plausible explanation, and the LNP points should therefore be able to be used as indicators of denitrification.

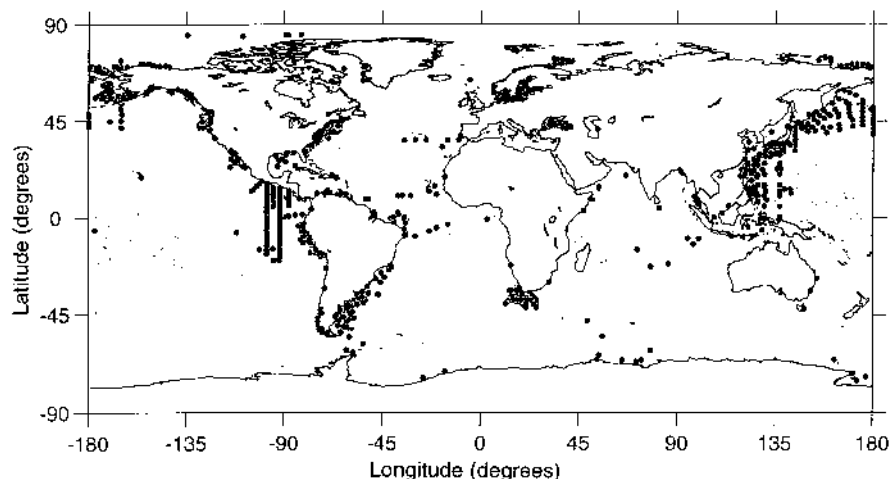


Figure 3 Geographical locations where (at any depth) measured phosphate concentration was $> 1.5 \mu\text{mol kg}^{-1}$ and the ratio of measured nitrate to measured phosphate concentrations was < 3.0 . (Data taken from Ref. 1). The distribution can be compared to a similar plot¹⁹ using an earlier and smaller NODC data set and using a different LNP criterion ($[\text{NO}_3^-] < ((15[\text{PO}_4^{3-}] - 1) - 10.0)$).

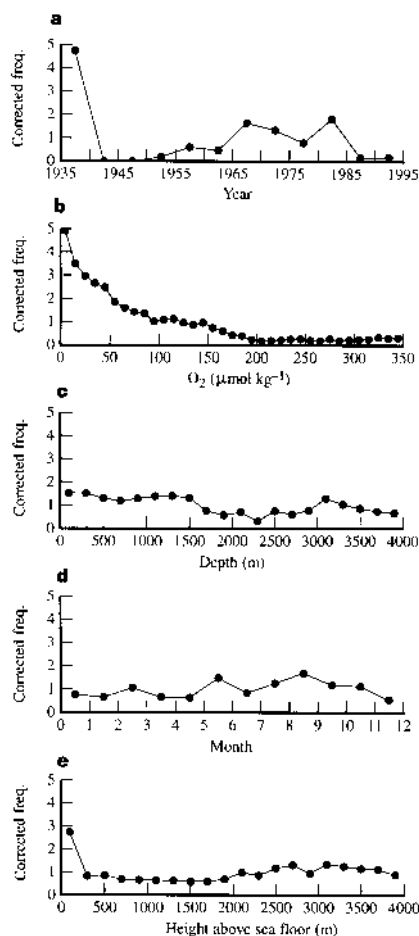


Figure 4 Frequency plots of the LNP points against: year of measurement (**a**); oxygen concentration (**b**); depth (**c**); month of the year (**d**); and height above sea floor (**e**). In each case the y-ordinate is calculated as $y = (\text{LNP}_x / \text{LNP}_{\text{tot}}) / (\text{NP}_x / \text{NP}_{\text{tot}})$. Here $(\text{LNP}_x / \text{LNP}_{\text{tot}})$ is the fraction of LNP points occurring at the value or range of the parameter plotted on the x-axis, and $(\text{NP}_x / \text{NP}_{\text{tot}})$ is the fraction of all points occurring at the same value or range. As an example, the y-ordinate of the left-most point in **b** is calculated as $(1,186 / 10,890) \div (9,742 / 613,169)$ because 10,890 of the LNP points and 613,169 of all the nitrate:phosphate points in the data set were measured at times when oxygen was also measured, and of those, 1,186 of the LNP points and 9,742 of the whole set of points were at oxygen concentrations of 0–20 $\mu\text{mol kg}^{-1}$. To ensure that the frequency plots are fully inter-comparable, the y-ordinates were then normalized so that the average y-ordinate for each plot is 1.0. The high correlation during 1935–40 in **a** is not significant as only 7 LNP points were measured during that period. The correlations in **e** are less reliable because sea-floor depths for each station are not available in ref. 1 and so had to be obtained instead by interpolation from values in a separate data set, at $(1/12) \times (1/12)$ degree resolution.

In those areas where denitrification is identified in this way (clusters of LNP points in Fig. 3) the denitrification may be occurring in the water column^{18,19} or within the sediment^{15,20}, with subsequent exchange of substrates and products with the overlying waters. Open-ocean denitrification has been reported in the eastern tropical North Pacific²¹, the eastern tropical South Pacific¹⁴, and in the Arabian Sea²². Our criteria (equation (2)) identify the eastern tropical Pacific (north and south), but not the Arabian Sea. The Arabian Sea does not appear as a denitrification site because, although denitrification takes place, the nitrate:phosphate ratio does not fall much below 9.0 as a result²². In light of this, it is clear that the method used here will detect the majority of denitrification-affected waters, but not all of them.

It has previously been suggested, from nitrogen deficit calculations, that denitrification is taking place within the northern North Pacific²³ and the Bering Sea²⁴, but nevertheless these areas are not traditionally recognized as major sites of denitrification. The calculations reported here provide additional evidence for denitrification in these places, and additionally across large areas of the western North Pacific (Fig. 3). Nitrate concentrations are generally high⁴ and oxygen concentrations low^{25,26} in sub-surface waters in these areas, providing favourable conditions for denitrification²⁷.

Assuming similar denitrification rates per unit area, the extent of the LNP points (Fig. 3) suggests that the total nitrate loss in the west and north North Pacific region is about 20% higher than in the eastern tropical Pacific Ocean. Sedimentary denitrification contributes about half as much as water-column denitrification to the global flux²⁰, and the eastern tropical Pacific (north and south) was previously estimated to contribute ~85% of total water-column denitrification²⁷. Therefore inclusion of the western and northern North Pacific could potentially increase the total global marine denitrification flux by about 70%. This calculation is speculative, but it is clear that more detailed work should be carried out to enable more precise estimates to be made. Denitrification is the largest sink of reactive nitrogen from the oceans^{28,29}, and is therefore a critical component in the global marine nitrogen cycle. □

Received 12 August 1996; accepted 21 April 1997.

1. NODC World Ocean Atlas 1994 data set (<http://www.nodc.noaa.gov/NODC-products.html>).
2. Geochemical Oceans Section Study (GEOSECS) data set (<http://ingrid.ligo.columbia.edu/SOURCES/GEOSECS/>).
3. Redfield, A. C. in *James Johnston Memorial Volume* 176–192 (Liverpool Univ. Press, 1934).
4. Levitus, S., Conkright, M. E., Reid, J. L., Najjar, R. G. & Mantyla, A. Distribution of nitrate, phosphate and silicate in the world oceans. *Prog. Oceanogr.* **31**, 245–273 (1993).
5. Anderson, L. A. On the hydrogen and oxygen content of marine phytoplankton. *Deep-Sea Res.* **14**, 1675–1680 (1995).
6. Copin-Montégut, C. & Copin-Montégut, G. Stoichiometry of carbon, nitrogen, and phosphorus in marine particulate matter. *Deep-Sea Res.* **30**, 31–46 (1983).
7. Smith, S. V. Phosphorus versus nitrogen limitation in the marine environment. *Limnol. Oceanogr.* **29**, 1149–1160 (1984).
8. Codispoti, L. A. in *Productivity of the Ocean: Present and Past* (eds Berger, W. H., Smetacek, V. S. & Wefer, G.) 377–394 (Wiley, New York, 1989).
9. Broecker, W. S. & Peng, T.-H. *Tracers in the Sea* (Eldigio, New York, 1982).
10. Fanning, K. A. Nutrient provinces in the sea—concentration ratios, reaction-rate ratios, and ideal covariation. *J. Geophys. Res.* **97**, 5693–5712 (1992).
11. Anderson, L. A. & Sarmiento, J. L. Redfield ratios of remineralization determined by nutrient data-analysis. *Glob. Biogeochem. Cycles* **8**, 65–80 (1994).
12. Meybeck, M. in *Interactions of C, N, P and S Biogeochemical Cycles and Global Change* (eds Wollast, R., Mackenzie, F. T. & Chou, L.) 163–193 (NATO ASI Ser., Vol. I 4, Springer, Berlin, 1993).
13. Kamykowski, D. & Zentara, S.-J. Hypoxia in the world ocean as recorded in the historical data set. *Deep-Sea Res.* **37**, 1861–1874 (1990).
14. Codispoti, L. A. & Christensen, J. P. Nitrification, denitrification and nitrous oxide cycling in the eastern tropical South Pacific Ocean. *Mar. Chem.* **16**, 277–300 (1985).
15. Canfield, D. E. in *Interactions of C, N, P and S Biogeochemical Cycles and Global Change* (eds Wollast, R., Mackenzie, F. T. & Chou, L.) 333–363 (NATO ASI Ser., Vol. I 4, 1993).
16. Allredge, A. L. & Silver, M. W. Characteristics, dynamics and significance of marine snow. *Prog. Oceanogr.* **20**, 41–82 (1988).
17. Kamykowski, D. & Zentara, S.-J. Spatio-temporal and process-oriented views of nitrite in the world ocean as recorded in the historical data set. *Deep-Sea Res.* **38**, 445–464 (1991).
18. Rönner, U. & Sörensson, E. Denitrification rates in the low-oxygen waters of the Baltic proper. *Appl. Environ. Microbiol.* **50**, 801–806 (1985).
19. Codispoti, L. A. et al. High nitrite levels off northern Peru: a signal of instability in the marine denitrification budget. *Science* **233**, 1200–1202 (1986).
20. Christensen, J. P., Townsend, D. W. & Montoya, J. P. Water column nutrients and sedimentary denitrification in the Gulf of Maine. *Continental Shelf Res.* **16**, 489–515 (1996).
21. Thomas, W. H. On denitrification in the northeastern tropical Pacific Ocean. *Deep-Sea Res.* **13**, 1109–1114 (1966).
22. Mantoura, R. F. C. et al. Nitrogen biogeochemical cycling in the northwestern Indian Ocean. *Deep-Sea Res.* **140**, 651–671 (1993).

23. Tsunogai, S. in *Biological Oceanography of the Northern North Pacific Ocean* (eds Takenouti, A. Y. et al.) 517–533 (Idemitsu Shoten, Tokyo, 1972).
24. Tsunogai, S., Kusakabe, M., Izumi, H., Koike, K. & Hattori, A. Hydrographic features of the deep water in the Bering Sea—the sea of silica. *Deep-Sea Res.* **26**, 641–659 (1979).
25. Levitus, S. & Boyer, T. P. *World Ocean Atlas 1994 Vol. 2, Oxygen* (Atlas NESDIS 2, NOAA, Washington DC, 1994).
26. Ogura, N. Relation between dissolved organic carbon and apparent oxygen utilization in the Western North Pacific. *Deep-Sea Res.* **17**, 221–231 (1970).
27. Hattori, A. in *Nitrogen in the Marine Environment* (eds Carpenter, E. & Capone, D.) 191–232 (Academic, New York, 1983).
28. Jaffe, D. A. in *Global Biogeochemical Cycles* (eds Butcher, S. S., Charlson, R. J., Orians, G. H. & Wolfe, G. V.) 263–284 (Academic, New York, 1992).
29. Mackenzie, F. T., Ver, L. M., Sabine, C., Lane, M. & Lerman, A. in *Interactions of C, N, P and S Biogeochemical Cycles and Global Change* (eds Wollast, R., Mackenzie, F. T. & Chou, L.) 1–61 (NATO ASI Ser., Vol. I 4, Springer, Berlin, 1993).
30. Conkright, M. E., Boyer, T. P. & Levitus, S. *Quality Control and Processing of Historical Nutrient Data* (Tech. Rep. NESDIS 79, NOAA, Washington DC, 1994).
31. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in FORTRAN 2nd edn* (Cambridge Univ. Press, 1992).

Acknowledgements. We thank D. Burton, C. Garside, D. Hydes and P. Wright for discussions; R. Mills, D. Hydes and M. Conkright for comments on the manuscript; and P. Challenor for sea-floor depth data.

Correspondence should be addressed to T.T. (e-mail: T.Tyrrell@soc.soton.ac.uk).

Contrasting physiological and structural vegetation feedbacks in climate change simulations

Richard A. Betts*, Peter M. Cox*, Susan E. Lee† & F. Ian Woodward†

* *Hadley Centre, Meteorological Office, Bracknell RG12 2SY, UK*

† *Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK*

Anthropogenic increases in the atmospheric concentration of carbon dioxide and other greenhouse gases are predicted to cause a warming of the global climate by modifying radiative forcing¹. Carbon dioxide concentration increases may make a further contribution to warming by inducing a physiological response of the global vegetation—a reduced stomatal conductance, which suppresses transpiration². Moreover, a CO₂-enriched atmosphere and the corresponding change in climate may also alter the density of vegetation cover, thus modifying the physical characteristics of the land surface to provide yet another climate feedback^{3–6}. But such feedbacks from changes in vegetation structure have not yet been incorporated into general circulation model predictions of future climate change. Here we use a general circulation model iteratively coupled to an equilibrium vegetation model to quantify the effects of both physiological and structural vegetation feedbacks on a doubled-CO₂ climate. On a global scale, changes in vegetation structure are found to partially offset physiological vegetation–climate feedbacks in the long term, but overall vegetation feedbacks provide significant regional-scale effects.

The Sheffield University vegetation model simulates global vegetation under steady-state conditions of climate and atmospheric CO₂ (ref. 7). It models the physiological processes of nutrient uptake, photosynthesis, respiration and stomatal limitation of transpiration, and uses these to determine the vegetation structural character in terms of foliage density. The outputs of this model are: (1) leaf area index (LAI), the area of leaf surface per unit area of ground; and (2) daytime mean canopy conductance (g_c), the net transpirational conductance of all stomata integrated (numerically) over the canopy depth. LAI is purely a structural variable, whereas g_c contains both structural and physiological contributions. The

contemporary vegetation simulation has been validated against point measurements⁷.

The Hadley Centre general circulation model (GCM) used here is a simplified version of that used for climate change prediction^{8,9}, consisting of an explicit representation of the global atmospheric circulation and a thermodynamic 'mixed-layer' ocean model with prescribed heat transports to represent ocean currents¹⁰. The simulations presented here neglect the relative cooling effect of increased sulphate aerosol concentrations⁸; this omission and that of explicit ocean current modelling means that the results cannot be regarded as a state-of-the-art prediction. Instead, they demonstrate the effect of vegetation feedback on climate sensitivity to atmospheric CO₂ concentrations.

The GCM land surface scheme is of moderate complexity¹¹, with the land surface state defined by seven prognostic variables: root-zone soil moisture, lying snow, intercepted canopy water, and the temperatures of four soil layers in the vertical. The surface energy partitioning, evapotranspiration, runoff and snowmelt are parametrized using driving variables from the atmosphere model and seven vegetation-specific land surface parameters³. The main parameters are: root depth, determining the depth of soil from which water can be extracted for transpiration; snow-free and deep-snow albedos, determining the fraction of incident solar radiation reflected from the surface; roughness length, determining the aerodynamic resistance for turbulent transfers; and surface conductance, determining the additional resistance for water vapour transfers in drought-free conditions. Over vegetated surfaces, the latter accounts for the control of transpiration by stomata, but is a prescribed vegetation-specific constant in this version of the scheme.

The GCM and vegetation model were coupled by iterating between the two models, each providing boundary conditions for the other. The GCM supplied climatological monthly means to the

vegetation model, which returned the global distributions of LAI and g_c . The latter were used to redefine the GCM land surface parameters for the next iteration, with surface conductance incorporating g_c directly, and the remaining structural parameters being derived semi-empirically from LAI (Fig. 1).

The physiological and structural vegetation feedbacks on CO₂-induced climate change were isolated and quantified using the following four coupled simulations. (1) Both climate and vegetation consistent with an atmospheric CO₂ concentration of 323 parts per million by volume, p.p.m.v. ($1 \times \text{CO}_2$; simulation C). (2) The climate at equilibrium under $2 \times \text{CO}_2$ (646 p.p.m.v.) radiative forcing, but with the physiological and structural characteristics of the vegetation held at the $1 \times \text{CO}_2$ state (simulation R). (3) $2 \times \text{CO}_2$ radiative forcing and $1 \times \text{CO}_2$ vegetation structure, but with surface conductance including direct effects of $2 \times \text{CO}_2$ and the associated climate change on plant physiology (simulation RP). (4) $2 \times \text{CO}_2$ radiative forcing with both the physiology and structure allowed to reach a new equilibrium state under $2 \times \text{CO}_2$ and the associated climate (simulation RPS).

The difference between simulations R and C represents the standard GCM sensitivity to CO₂ excluding vegetation feedbacks, and the difference between RP and R defines the additional climate change resulting from the direct physiological effects (a comparable experiment to that in ref. 2). Finally, the difference between RPS and R demonstrates the combined effect of physiological and structural vegetation change on the climate sensitivity; this is the main new result of this work.

The radiation-only $2 \times \text{CO}_2$ sensitivity (R – C) of this version of the GCM was 4.3 K, which is at the high end of the IPCC range¹. The modelled climate change showed relatively large changes in temperature and precipitation in the tropics (Fig. 2), associated with strong cloud-mediated feedbacks. The physiological response in

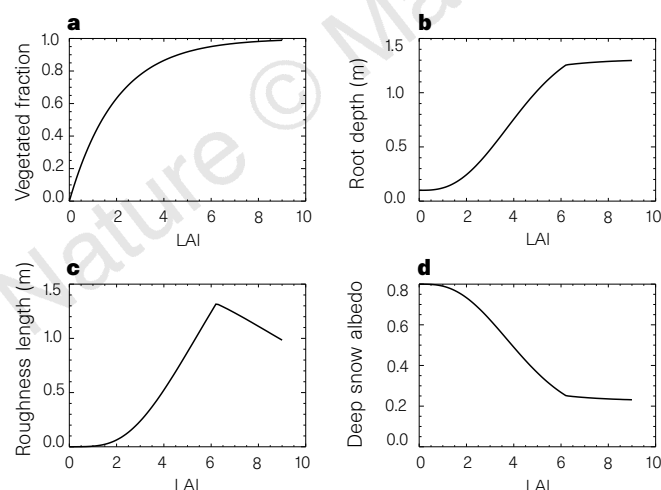


Figure 1 Derivation of GCM land surface parameters from leaf area index (LAI). **a**, Vegetated fraction is the fractional ground area covered by vegetation; this determines the relative weighting of values appropriate to vegetation and soil for the other variables. An empirical fit with the literature data^{16–18} shows that vegetated fraction increases with LAI, saturating at higher values. **b**, The mean root depth is found empirically^{16–20} to increase with LAI in the grid-box mean, and is also assumed to be linearly correlated with the maximum infiltration rate at the soil surface. **c**, Roughness length increases with LAI at low leaf areas, but decreases with LAI at higher values as the canopy becomes closed²¹. **d**, Deep-snow albedo¹⁷ is the upper limit of surface reflectance in deep-snow conditions, and this decreases with LAI as more of the snow is masked by the darker vegetation. Snow-free albedo shows a similar but less pronounced relationship.

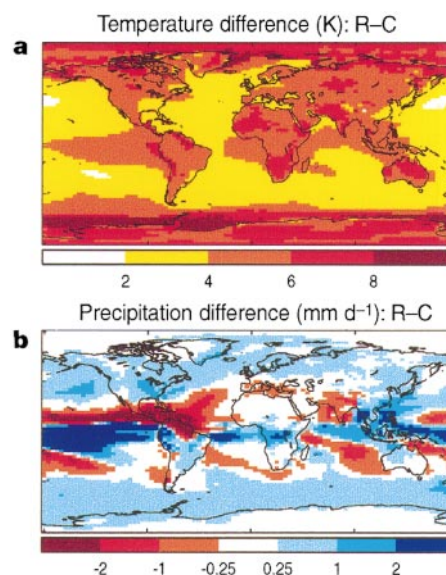


Figure 2 Climate change due to doubling the atmospheric concentration of CO₂, neglecting vegetation feedback, expressed as differences between simulations R and C (see text). **a**, Change in annual mean temperature, diagnosed at a height of 1.5 m above the surface. **b**, Change in annual mean precipitation.

Key to simulations

R : $2\times\text{CO}_2$ radiation
 RP : $2\times\text{CO}_2$ radiation + physiological feedback
 RPS : $2\times\text{CO}_2$ radiation + physiological & structural feedback

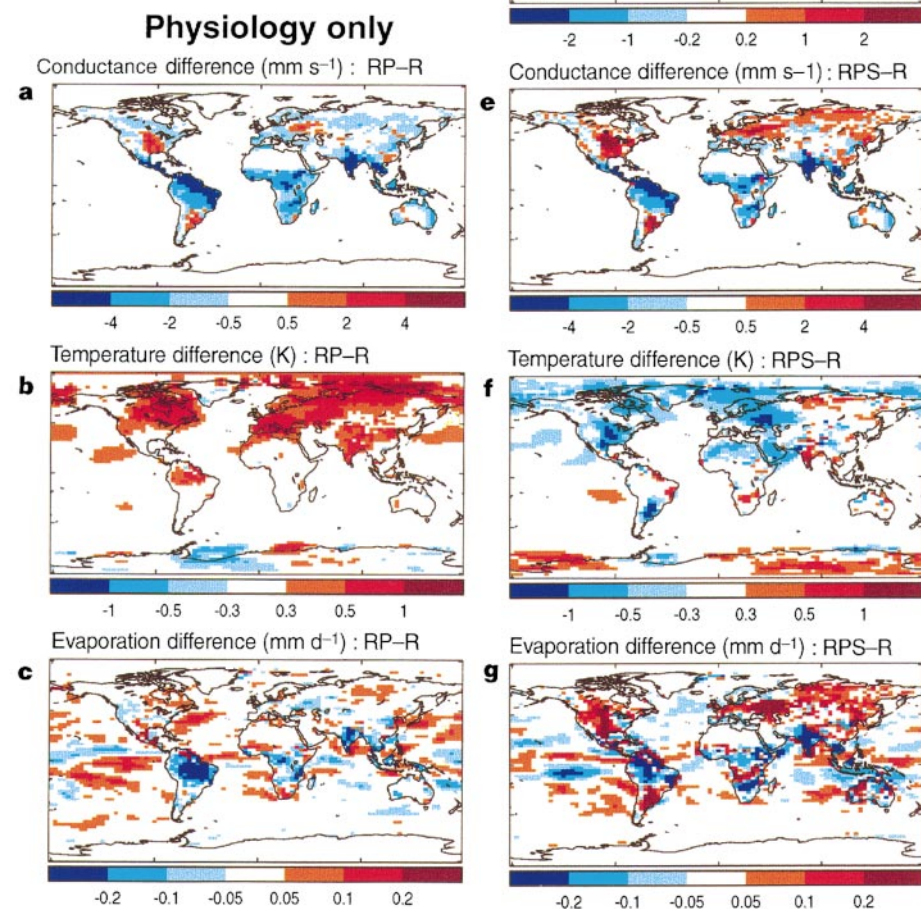


Figure 3 Physiological and structural vegetation change under doubled atmospheric CO_2 concentration ($2\times\text{CO}_2$), and feedback of each on $2\times\text{CO}_2$ climate. **a**, Change in canopy conductance due to physiological response to $2\times\text{CO}_2$ and the associated climate change. **b**, Effect of physiological feedback on $2\times\text{CO}_2$ temperature. **c**, Effect of physiological feedback on $2\times\text{CO}_2$ evaporation. **d**, Change in leaf area index due to structural response to $2\times\text{CO}_2$ and the associated climate change. **e**, Change in canopy conductance due to both physiological and structural response. **f**, Combined effect of both physiological and structural feedback on $2\times\text{CO}_2$ temperature. **g**, Combined effect of both physiological and structural feedback on $2\times\text{CO}_2$ evaporation. 40% of the land surface experienced vegetation feedbacks on temperature (RPS - R) of at least 5% of the magnitude of the changes due to radiative forcing alone (R - C). 13% of the land showed relative temperature feedbacks of 10% or more. The relative evaporation feedbacks were larger and more widespread, with 74% of the land surface having a relative feedback of 10% or more, and 30% showing a feedback of over 50%. Vegetation-induced evaporation changes were larger than the greenhouse-gas-only changes for 18% of the land surface. Calculation of *t*-statistics for grid-point annual means showed that most temperature changes of 0.5 K or more were significant at the 1% confidence level or better. The exceptions to this were in the polar and sub-polar regions, where significance was reduced by higher interannual variability. In Siberia, temperature changes of 0.5 K were significant at 5% or better, while in Antarctica and the Arctic Ocean, little of the temperature change was significant at better than 20%. Almost all land evaporation changes of 0.1 mm d^{-1} were significant at 1%.

simulation RP was a general reduction in g_c relative to simulation R (Fig. 3a), consistent with increased water-use efficiency under $2\times\text{CO}_2$. Some areas with modified hydrological regimes experienced g_c increases caused by increased humidity, but the global mean change was a reduction of $\sim 20\%$ (Table 1). These caused significant feedbacks on climate (RP - R), with temperature increasing by up to 1 K over Northern Hemisphere land (Fig. 3b). The large conductance decreases in the tropical forests produced small temperature changes but appreciable reductions in evapotranspiration (Fig. 3c). The modelled effects of physiology on mean land temperature, evapotranspiration and conductance are all in close agreement with those from a previous study².

The structural response in simulation RPS was a widespread increase in LAI relative to simulation R (Fig. 3d), due to increased productivity and water-use efficiency under the new CO_2 concentration and climate. The greatest LAI increases were in regions of increased rainfall. These changes acted to offset the physiological reductions in conductance, and at high latitudes the result was an overall increase in g_c (Fig. 3e); this is contrary to the result obtained when allowing physiological change alone (Fig. 3a). Elsewhere, the reduced g_c seen in simulation RP also occurred in simulation RPS;

the reductions were smaller than in RP, except in regions where significantly reduced rainfall (Fig. 2b) had caused conspicuous reductions in LAI (Fig. 3d). The combined effect of physiology and structure was a reduction of $\sim 12\%$ in g_c in the global mean (Table 1), which is considerably less than the reduction due to physiology alone.

The combined physiological and structural vegetation feedbacks had significant effects on the climate sensitivity (RPS - R; Fig. 3f,g). Structural changes acted via two competing effects; increased LAI tended to warm the land surface by lowering its albedo⁴⁻⁶ and to cool the land surface by enhancing evaporation (and consequently cloud cover^{12,13}) via increases in root depth¹⁴, roughness length¹⁵ and surface conductance. Similarly, decreased LAI tended to cool the surface via increased albedo, and warm the surface via reduced evaporation. The albedo effect dominated in regions where the vegetation was sparse, or where the underlying surface was much more reflective than the vegetation such as in snow-covered regions⁴⁻⁶; however, the evaporation effect dominated elsewhere. Temperature changes (Fig. 3f) were therefore negatively correlated with LAI changes (Fig. 3d), except in sparsely vegetated regions and also northern Siberia, where greater LAI caused a warming via

Table 1 Global mean vegetation feedbacks on $2 \times \text{CO}_2$ climate

Variable	R	RP – R	RPS – R
	Mean over land		
LAI	3.38	0.0%	7.2%
g_c	6.09 mms^{-1}	– 19.6%	– 12.1%
T_s	284.4 K	0.2 K	– 0.1 K
P	2.50 mm d^{-1}	– 0.7%	– 0.2%
E	1.54 mm d^{-1}	– 1.8%	– 0.3%
	Mean over land and ocean		
T_s	290.2 K	0.2 K	– 0.1 K
P	3.41 mm d^{-1}	0.0%	– 0.1%
E	3.41 mm d^{-1}	0.0%	– 0.1%

Mean values and perturbations to leaf area index (LAI), canopy conductance (g_c), screen-level temperature (T_s), total precipitation (P) and total surface moisture flux (E). Column 2 shows absolute values for simulation R (radiation only), column 3 shows the changes due to the physiological response only (RP – R), and column 4 gives the total vegetation feedback (RPS – R). Changes are given as percentages of the value in simulation R, except for temperature changes which are given in K.

increased masking of snow. The feedback through evaporation was significantly modified by structural changes, especially in the middle- and high-latitude regions (compare Fig. 3c and g). However, transpiration from the tropical rainforests, which experienced negligible changes in LAI, was still significantly reduced compared to simulation R.

It is important to recognize that changes in vegetation structure may lag the physiological response to increased CO_2 by several years or even decades. Therefore, the actual effect of vegetation feedback on climate at the time of CO_2 doubling is likely to lie somewhere between the results of simulations RP and RPS. A full assessment of this will require a model of vegetation dynamics fully integrated within a GCM. Nevertheless, our results show that changes in land surface properties due to vegetation can provide climatic feedback mechanisms that are both positive and negative in relation to climate change due to radiative forcing alone; furthermore, they demonstrate that the sign of the feedback depends partly on whether local vegetation growth is enhanced or suppressed by increased CO_2 concentration and the associated climate change, and partly on the nature of the locally dominant surface–atmosphere interaction. Both physiological and structural characteristics of the vegetation have been shown to be important, with changes in one property often counteracting changes in another. In the global mean, the competing effects of increased water use efficiency and increased LAI cause a small surface evaporation change relative to the climate change simulation with fixed vegetation properties. We conclude that a short-term enhancement of regional climate warming by vegetation physiology may eventually be mitigated by a longer term modification of surface characteristics due to vegetation morphology. As this work does not account for the timescales involved in the full suite of vegetation feedbacks, the next stage should be to include dynamical changes in both vegetation physiology and structure in GCM predictions of future climate change. □

Received 28 October 1996; accepted 28 April 1997.

- Houghton, J. T. *et al.* (eds) *Climate Change 1995* (Cambridge Univ. Press, 1995).
- Sellers, P. J. *et al.* Comparison of radiative and physiological effects of doubled atmospheric CO_2 on climate. *Science* **271**, 1402–1406 (1996).
- Lean, J. & Rowntree, P. R. A GCM simulation of the impact of Amazonian deforestation on climate using an improved canopy representation. *Q. J. R. Meteorol. Soc.* **119**, 509–530 (1993).
- Bonan, G. B., Pollard, D. & Thompson, S. L. Effects of boreal forest vegetation on global climate. *Nature* **359**, 716–718 (1992).
- Foley, J. A., Kutzbach, J. E., Coe, M. T. & Levis, S. Feedbacks between climate and boreal forests during the Holocene epoch. *Nature* **371**, 52–54 (1994).
- Gallimore, R. G. & Kutzbach, J. E. Role of orbitally induced changes in tundra area in the onset of glaciation. *Nature* **381**, 503–505 (1996).
- Woodward, F. I., Smith, T. M. & Emanuel, W. R. A global land primary productivity and phytoecography model. *Glob. Biogeochem. Cycles* **9**, 471–490 (1995).

- Mitchell, J. F. B., Johns, T. C., Gregory, J. M. & Tett, S. F. B. Climate response to increasing levels of greenhouse gases and sulphate aerosols. *Nature* **376**, 501–504 (1995).
- Jones, R. G., Murphy, J. M. & Noguer, M. Simulation of climate change over Europe using a nested regional-climate model. I: Assessment of control climate, including sensitivity to location of lateral boundaries. *Q. J. R. Meteorol. Soc.* **121**, 1413–1449 (1995).
- Senior, C. A. & Mitchell, J. F. B. Carbon dioxide and climate: the impact of cloud parameterization. *J. Clim.* **6**, 393–418 (1993).
- Chen, T. H. *et al.* Cabauw experimental results from the project for intercomparison of land-surface schemes (PILPS). *J. Clim.* (in the press).
- Rowntree, P. R. & Bolton, J. A. Simulation of the atmospheric response to soil moisture anomalies over Europe. *Q. J. R. Meteorol. Soc.* **109**, 501–526 (1983).
- Beljaars, A. C. M., Viterbo, P., Miller, M. J. & Betts, A. K. The anomalous rainfall over the United States during July 1993: sensitivity to land surface parameterization and soil moisture anomalies. *Mon. Weath. Rev.* **124**, 362–383 (1996).
- Milly, P. C. D. & Dunne, K. A. Sensitivity of the global water cycle to the water-holding capacity of land. *J. Clim.* **7**, 506–526 (1994).
- Rowntree, P. R. in *Land Surface Evaporation* (eds Schmugge, T. J. & Andre, J. C.) 5–30 (Springer, New York, 1991).
- Wilson, M. F. & Henderson-Sellers, A. A global archive of land cover and soils data for use in general circulation climate models. *J. Climatol.* **5**, 119–143 (1985).
- Sellers, P. J., Mitz, Y., Sud, Y. C. & Dalcher, A. A simple biosphere model (SiB) for use within general circulation models. *J. Atmos. Sci.* **43**, 505–531 (1986).
- Lowry, W. P. & Lowry, P. P. *Fundamentals of Biometeorology. Interactions of Organisms and the Atmosphere* Vol. 1 (Peavine, McMinnville, OR, 1989).
- Eagleson, P. S. *Dynamic Hydrology* (McGraw Hill, New York, 1970).
- Halldin, S., Saugier, B. & Pontallier, F. Y. Evapotranspiration of a deciduous forest. Simulation using routine meteorological data. *J. Hydrol.* **75**, 323–341 (1984).
- Shaw, R. H. & Pereira, A. R. Aerodynamic roughness of a plant canopy. *Agricult. Meteorol.* **26**, 51–65 (1982).

Acknowledgements. We thank E. M. Blyth, J. Foley, R. J. Harding, W. J. Ingram, J. E. Lovelock, J. F. B. Mitchell, P. L. Mitchell, P. R. Rowntree, C. A. Senior, W. J. Shuttleworth, S. F. B. Tett and P. J. Valdes for comments, advice and discussion. This work was supported by the NERC TIGER programme and the UK Department of the Environment.

Correspondence and requests for materials should be addressed to R.A.B. (e-mail: rabetts@meto.gov.uk).

Wing upstroke and the evolution of flapping flight

Samuel O. Poore, A. Sánchez-Haiman & G. E. Goslow Jr

Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA

Movements of the wing during upstroke in birds capable of powered flight are more complex than those of downstroke^{1–3}. The m. supracoracoideus (SC) is a muscle with a highly derived morphology that is generally considered to be the primary elevator of the wing^{4–6}. This muscle arises from the ventrally oriented sternum and its tendon of insertion passes craniodorsally through a special bony canal, around a bony process which deflects it laterally, to attach on the dorsal aspect of the humerus above the glenohumeral joint (Fig. 1). We studied the contractile properties of the SC *in situ* and related them to wing kinematics in the European starling (*Sturnus vulgaris*). Our findings indicate that the primary role of the SC is to impart a high-velocity rotation about the longitudinal axis of the humerus. This rapid ‘twisting’ of the humerus, coupled with limited humeral elevation, is responsible for positioning the forearm and hand so that their subsequent extension orients the outstretched wing appropriately for the following downstroke. This reinterpretation of the primary function of the SC provides insight into the selective advantage of its unique musculoskeletal organization in the evolution of powered flapping flight in birds.

A general feature of powered locomotion based on an oscillating wing is an asymmetry in how the wing meets the environment during the downstroke compared with the upstroke parts of the wingbeat cycle⁷. The downstroke in birds, when primary lift and propulsion are achieved, is characterized by an outstretched wing. The more complicated upstroke involves rapid withdrawal of the wing towards the body to reduce its surface area, elevation and subsequent extension in a way that minimally retards lift and thrust gained in the previous downstroke. The distinct musculoskeletal configuration of the m. supracoracoideus (SC) was not present in

the Jurassic bird *Archaeopteryx*^{8,9}, nor is there firm evidence for its presence in recently described Late Jurassic or Early Cretaceous species^{10–15}. The evolution of this condition is thought to have been important for flapping powered flight in improving the function of wing elevation^{16,17}, but the significance of its highly derived organization has not been appreciated.

We measured isometric forces of rotation (torque) about the longitudinal axis of the humerus, the extent of unrestrained humeral elevation and rotation, and the muscle's contractile properties, by direct nerve stimulation of the SC *in situ*. Stimulation of the SC at joint angles of elevation/depression and protraction/retraction coincident with the downstroke–upstroke transition and mid-upstroke² imparted a substantial mean isometric torque about the longitudinal axis of the humerus (Fig. 2). A mean isometric value of 4.9 N ($n = 2$) was recorded at the downstroke–upstroke transition. In the humeral excursion experiments, we measured *in situ* humeral rotations of up to 70° and maximum elevations of 50°. In three length–force experiments, we measured a mean ($\pm$ s.d.) maximal tetanic force of 6.5 ($\pm$ 1.2) N, approximately 10 times body weight. The length of the SC coincident with the downstroke–upstroke transition corresponds to a position on the ascending limb of the active length–force curve of 3.0 N (Fig. 3). In lateral view, the right humerus rotates anticlockwise about its longitudinal axis and elevates a total of 50° during upstroke (-10° to 40°); these movements correspond to muscle shortening of 3.0 mm and a coincident decrease in the SC's potential for active force production (Fig. 3). The ascending arm of the active length–force curve occurs over a short distance (5 mm) and is steep owing to the short fascicle lengths associated with the SC's bipinnate architecture. These data reveal that, although the SC is capable of elevation of the humerus, a far more important role is to provide a high-velocity rotation about its longitudinal axis.

Further support for this conclusion can be drawn from the mechanical organization of the SC, its electrical activity patterns reported during slow and fast flight, and the highly derived morphology of the avian shoulder. The bipinnate structure of the SC, with its relatively short but numerous fascicles characteristic of all birds we examined, is an architecture for production of high

force. The SC's moment arm for humeral rotation about the longitudinal axis is short; we estimate its maximum in the starling to be 2.0 mm. When a large torque (as provided by the SC) is applied across this short moment, the resultant velocity of humeral rotation and its attached distal wing element is magnified.

Aerodynamic models predict that active muscle contractile force is far more important for wing elevation in birds during slow than in fast flapping flight¹⁸. Passive aerodynamic lift is sufficient at fast flight speeds, particularly in birds with high-aspect ratio wings. In four species of birds with diverse flight styles for which electromyograms of the SC during flight have been reported, however, the muscle remains electrically active over a range of flight speeds^{2,4,19,20}. Passive aerodynamic lift may not be implemented at the flight speeds attained in these experimental situations. The SC may act to adjust the planform of the wing's leading edge at all speeds. Electromyographic activity does not necessarily correlate with useful mechanical force²¹ and its presence may simply reflect out-flow modulation of a central neural control program²². Or, as we believe, the activity of the SC at even high flight speeds is necessary for its most important role, that of humeral rotation.

The avian shoulder joint is structurally derived and functionally complex. The glenoid, best described as a hemisellar (half saddle) joint, is concavoconvex in configuration, faces dorsolaterally, and articulates with a bulbous humeral head. Jenkins²³ reviewed the evolution of this joint in a comparative study and provided a new interpretation of its functional morphology based on a cineradiographic analysis of the wingbeat cycle. The articulation of the humeral head on the dorsally facing surface of the glenoid (the *Labrum cavitatis glenoidalis*) at the upstroke–downstroke transition allows full abduction of the wing into the parasagittal plane. Our observations further reveal that this extensive abduction of the wing, characteristic of so many flying birds, is accomplished not by simple elevation of the humerus, but primarily by rotation about its longitudinal axis.

In a classic study of the functional organization of the wing²⁴, Sy described humeral axial rotation as a mechanism for the execution of wing upstroke in pigeons and generalized its importance for other relatively small birds with powered flight. To appreciate the

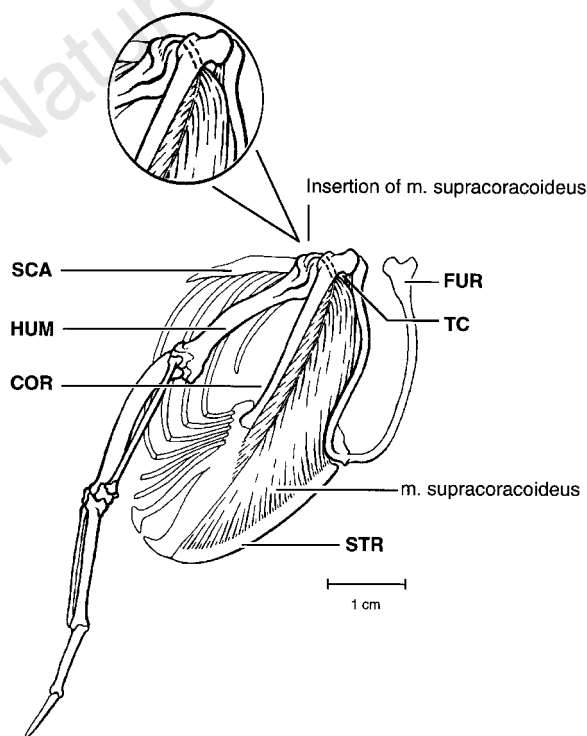


Figure 1 Anterolateral view of the right shoulder in the European starling (*Sturnus vulgaris*). The m. pectoralis has been removed, as well as all other wing and shoulder musculature to expose the m. supracoracoideus (SC) and its tendon of insertion on the dorsal aspect of the humerus. The fascicles of the SC arise from the dorsal half of the carina, the adjacent body of the sternum, and a small area on the base of the coracoclavicular membrane. The SC's bipinnate architecture limits tendon excursion but maximizes force production. Scale bar, 1.0 cm. Abbreviations: SCA, scapula; HUM, humerus; COR, coracoid; TC, triosseal canal; FUR, furcula, STR, sternum.

functional significance of this rotation, it is necessary to consider the orientation of the humerus to the body axis. When viewed from above (in dorsal view), the angle formed by the long axis of the humerus and the bird's longitudinal axis when the wing is in the downstroke-upstroke transition position is not 90° as might be expected: the humerus is retracted back towards the body to form an acute angle with the long axis of the back. At the downstroke-upstroke transition in starlings, for example, this angle is at its maximum of $55\text{--}60^\circ$ during early upstroke, and decreases to its minimum of $25\text{--}30^\circ$ (that is, the humerus is retracted). Simple elevation of a retracted humerus orients the wing's ventral surface posterolaterally instead of in the more functional lateral position. Simultaneous rotation and elevation of the humerus in early upstroke is necessary so that subsequent extension of the hand and forearm during late upstroke will orient the fully outstretched

wing in the parasagittal plane; that is, the wing's ventral surface will face laterally, the position appropriate for the beginning of the subsequent downstroke. Humeral rotation/retraction as described here is key to the execution of high-amplitude, high-frequency wingbeats in the European starling, a generalized member of the order Passeriformes, and may also be important in other birds, including those with longer wings, as well as for other flying vertebrates. For example, in a study of the shoulder of the Cretaceous pterodactyloid pterosaur *Santanadactylus brasiliensis*, which has an estimated body mass of $3.9\text{--}7.3$ kg and a total wingspan of 4.7 m, humeral rotation was proposed to be the key to wing elevation in this species²⁵.

In the seven preserved specimens of the Jurassic bird *Archaeopteryx*, there is no evidence of a derived SC with a dorsally inserting tendon, as indicated by the lack of a triosseal canal or an

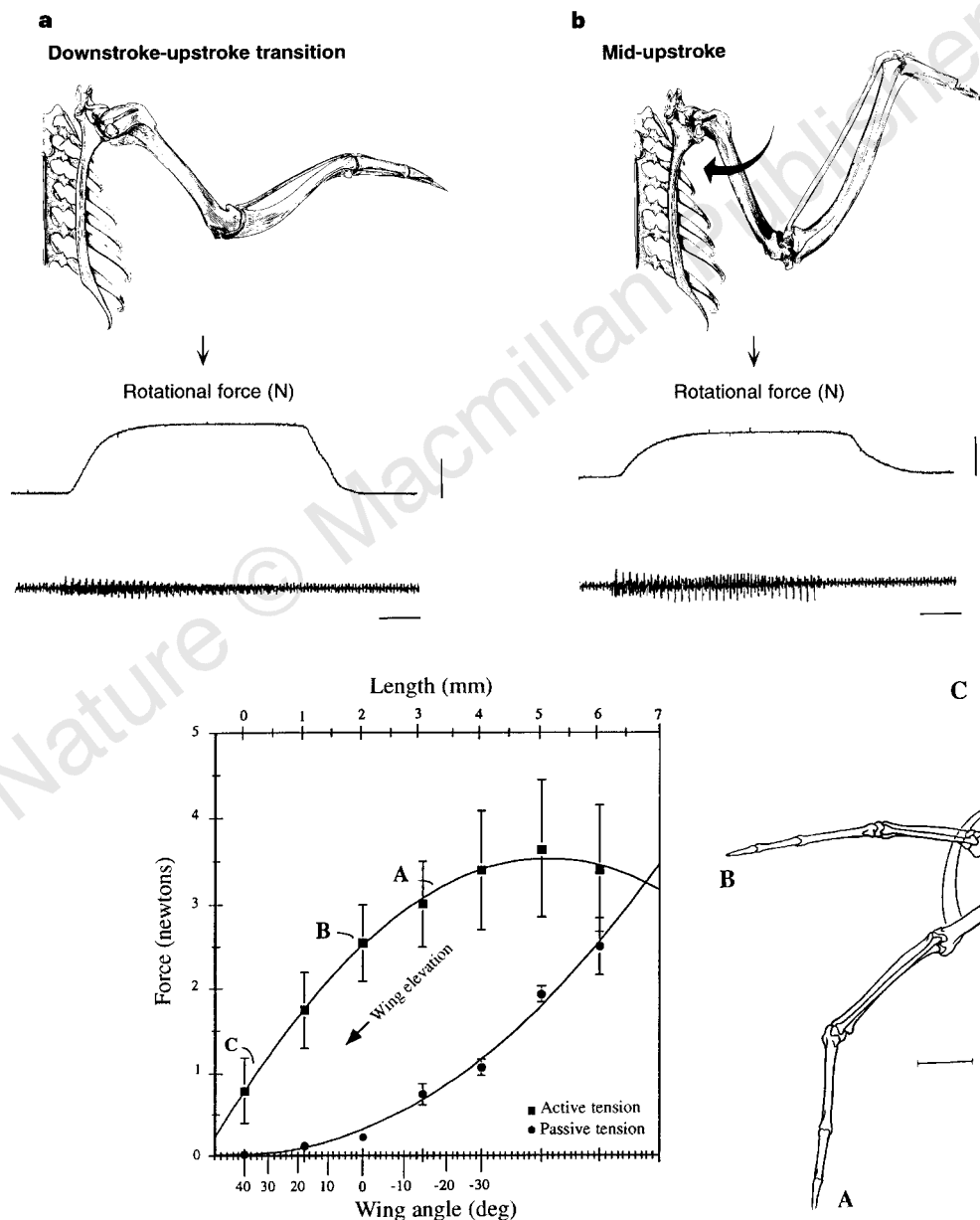


Figure 3 Functional correlates of length-force in European starlings (*Sturnus vulgaris*). Top, active and passive length-force characteristics of the supracoracoideus muscle derived by direct nerve stimulation from four *in situ* experiments. Bottom, anterior views of the right shoulder at the downstroke-upstroke transition (A), mid-upstroke (B) and late upstroke (C). The humerus is

Figure 2 Rotational force of the m. supracoracoideus (SC) measured at the deltopectoral crest. We measured *in situ* isometric rotational force (torque) produced by the SC in two experiments at two positions of the humerus relevant to the wingbeat cycle; downstroke-upstroke transition (a) and mid-upstroke (b). Force measured at the downstroke-upstroke transition (a) was consistently higher than that at mid-upstroke (b). The moment arm of the SC for rotation is short (2 mm). The resulting mechanical advantage for rotation of the humerus by the SC is relatively low, but when input forces are high, the arrangement is favourable for the production of high-velocity rotation at the distal wing. Time bar, 100 ms; force bar, 2.5 N.

depressed 10° below the horizontal at the downstroke-upstroke transition, which corresponds to a position on the ascending arm of the active length-force curve (A) where the muscle is capable of producing 85% of total force. As the wing is elevated/rotated to the mid-upstroke (B) and late-upstroke (C) positions, the capacity for force production by the SC declines. Scale bar, 1.0 cm.

acrocoracoid^{18,9,26,27}. The lack in *Archaeopteryx* of this organization, and therefore of rapid humeral rotation, may have restricted its ability to execute rapidly the high-amplitude wing movements associated with slow flight, take-off and landing in modern species of similar size. The early fossil record of birds since the Jurassic demonstrates a number of advanced characteristics that signal evolutionary improvements towards powered flight. For example, the glenoid of *Archaeopteryx* and the glenoid of the younger Lower Cretaceous *Sinornis* both face laterally^{8,9,23}, which presumably limited extensive abduction of the wing into the parasagittal plane in both species. In the recently described series of Enantiornithine and non-Enantiornithine Mesozoic birds (that is, *Eoalulavis*, *Neuquenornis*, *Cathayornis*, *Concornis*, *Ambiortis* and *Iberomesornis*)^{10,12,15,28–30}, the presence of an elongated coracoid, furcula and scapular acromion suggests that a derived SC may have been present, although an acrocoracoid or triosseal canal has not been identified *de facto*.

Our reinterpretation here of the primary action of the SC, that of high-velocity humeral rotation, explains a new aspect of the evolutionary significance of its highly derived morphology in birds. The selective advantage of high-velocity rotation of the humerus to position the wing appropriately for downstroke was undoubtedly an important advancement to powered flight which contributed to the extensive adaptive radiation witnessed in modern birds. □

Methods

Forces of rotation and elevation. Under deep anaesthesia (ketamine 60 mg per kg, xylazine 6 mg per kg, with supplemental ketamine given as needed), we bisected the latissimus dorsi and rhomboideus muscles to expose the brachial plexus and isolate the nerve to the SC. A trachea tube was inserted to provide unidirectional ventilation (80% O₂, 20% N₂) after opening the posterior air sacs. All other components of the brachial plexus were severed to prevent stimulation of adjacent muscles, and the nerve to the SC was mounted on silver bipolar electrodes. We stabilized the bird by clamping the sternal keel, proximal coracoid and vertebral margin of the scapula to a heavy frame and body temperature was maintained at 39°C with warmed avian ringers and a heat lamp. A silver wire (0.38 mm diameter) was threaded through a small hole drilled in the deltopectoral crest and attached to a Grass T4 force transducer. The humerus was stabilized by inserting a needle (23 gauge) into its distal shaft, thus allowing 'free' rotation about the long axis of the humerus while restricting elevation. At wing positions corresponding to the downstroke–upstroke transition and mid-upstroke, a supramaximal stimulus (0.2-ms pulse, 60 Hz, 500-ms train) was delivered to the nerve of the SC. Substantial torque was produced by the SC around the longitudinal axis of the humerus in both positions. After removing the needle from the shaft of the humerus, we measured forces of humeral elevation by securing a piece of 5–0 surgical silk (compliance, 0.45 µm N⁻¹ cm⁻¹) to the humerus at midshaft. This configuration detected the elevational component of force but allowed the humerus to rotate on its long axis. Forces of elevation were consistently lower than forces of rotation.

Excursion of the humerus. We measured the total *in situ* rotational excursions of the humerus during tetanus of the SC for two starlings. We measured humeral rotation by placing a 23-gauge pin guided by a rack and pinion through a small hole drilled in the distal head of the humerus. This pin served as a pivot for rotation while restricting the elevational component of movement. We placed a 26-gauge pin perpendicular to the long axis of the humerus, which served as a dial with which to measure the degree of rotation. We removed the 23-gauge pin and allowed the humerus to move during stimulation to measure humeral elevation. We stimulated the nerve tetanically (60 Hz; 500-ms train duration) and measured humeral rotation with a protractor.

Active and passive length–force. We measured the muscle's passive and active properties at a series of lengths (Fig. 3, top; abscissa) within the muscle's normal *in vivo* excursion. We established the active length–force curves from maximal twitch responses (single stimulus, 0.2-ms pulse) and measured maximum whole-muscle tetanic force (60 Hz, 500-ms train) at the end of each experiment at the length coincident with maximum twitch force. After isolating the nerve to the SC (as already described), we deflected the deltoid and

propatagialis muscles from the dorsal aspect of the humerus, cut the bone around the tendon's site of attachment, and removed the bone chip together with the tendon. With the bird stabilized to a heavy frame by clamping the sternal keel, proximal coracoid and vertebral margin of the scapula, the bone/tendon structure was secured to the force transducer with a silk tie (compliance, 0.45 µm N⁻¹ cm⁻¹). A force transducer (Grass FT4) was mounted on an adjustable rack and pinion (millimetre calibration), which allowed us to change the muscle's length in millimetre increments. We generated passive length–force curves by lengthening the muscle in 1.0-mm increments over its physiological working range. We returned the muscle to the length of 'zero' passive force after each measurement before pulling it to a new length. The active length–force curve was generated by following the same length-change protocol and delivering a single supramaximal stimulus (0.2 ms) to the nerve at each length. We correlated absolute muscle length to wing angle of elevation and depression (Fig. 3, bottom; abscissa) by manipulating the contralateral wing in the bird after killing (sodium pentobarbital, 100 mg kg⁻¹). We dissected the origin of the SC from its attachment on the coracoclavicular membrane and sternum and attached it to a rack and pinion. This configuration allowed us to manipulate wing elevation, depression and rotation, and to correlate these with changes in tendon excursion. We manipulated the humerus in 10° increments above and below the horizontal. □

Received 2 January; accepted 7 April 1997.

1. Brown, R. H. J. Flapping flight. *Ibis* **93**, 333–359 (1951).
2. Dial, K. P., Goslow, G. E. Jr & Jenkins, F. A. Jr The functional anatomy of the shoulder in the European starling (*Sturnus vulgaris*). *J. Morphol.* **207**, 327–344 (1991).
3. Simpson, S. F. The flight mechanism of the pigeon *Columba livia* during take-off. *J. Zool.* **200**, 435–443 (1983).
4. Dial, K. P., Kaplan, S. R., Goslow, G. E. Jr & Jenkins, F. A. Jr A functional analysis of the primary upstroke and downstroke muscles in the domestic pigeon (*Columba livia*) during flight. *J. Exp. Biol.* **134**, 1–16 (1988).
5. Norberg, U. M. *Vertebrate Flight* 1–291 (Springer, Berlin, 1989).
6. Rayner, J. M. V. The evolution of vertebrate flight. *Biol. J. Linn. Soc.* **34**, 269–287 (1988).
7. Spedding, G. R., Rayner, J. M. V. & Pennycuik, C. J. Momentum and energy in the wake of a pigeon (*Columba livia*) in slow flight. *J. Exp. Biol.* **111**, 81–102 (1984).
8. Ostrom, J. H. *Archaeopteryx* and the origin of birds. *Biol. J. Linn. Soc.* **8**, 91–182 (1976).
9. Ostrom, J. H. Some hypothetical anatomical stages in the evolution of avian flight. *Smithson. Contr. Paleobiol.* **27**, 1–21 (1976).
10. Sanz, J. L. et al. An Early Cretaceous bird from Spain and its implications for the evolution of avian flight. *Nature* **382**, 442–445 (1996).
11. Hou, L., Zhou, Z., Martin, L. D. & Feduccia, A. A beaked bird from the Jurassic of China. *Nature* **377**, 616–618 (1995).
12. Chiappe, L. M. & Calvo, J. O. *Neuquenornis volans*, a new Late Cretaceous bird (Enantiornithes: Avisuridae) from Patagonia, Argentina. *J. Vert. Paleol.* **14**, 230–246 (1994).
13. Hou, L. A late mesozoic bird from inner Mongolia. *Vert. Pal. Asiatica* **32**, 258–266 (1994).
14. Sereno, P. C. & Rao, C. Early evolution of avian flight and perching: New evidence from the Lower Cretaceous of China. *Science* **255**, 845–848 (1992).
15. Zhou, Z., Jin, F. & Zhang, J. Preliminary report on a Mesozoic bird from Liaoning, China. *Chinese Sci. Bull.* **37** (16), 1365–1368 (1992).
16. Olson, S. L. & Feduccia, A. Flight capability and the pectoral girdle of *Archaeopteryx*. *Nature* **278**, 247–248 (1979).
17. Rayner, J. M. V. in *Biomechanics in Evolution* (eds Rayner, J. M. V. & Wootton, R. J.) 183–212 (Cambridge Univ. Press, 1991).
18. Rayner, J. M. V. Form and function in avian flight. *Curr. Ornithol.* **5**, 1–77 (1988).
19. Meyers, R. A. Gliding flight in the American kestrel (*Falco sparverius*): An electromyographic study. *J. Morphol.* **215**, 213–224 (1993).
20. Tobalske, B. W. & Dial, K. P. Neuromuscular control and kinematics of intermittent flight in budgerigars (*Melospittacus undulatus*). *J. Exp. Biol.* **187**, 1–18 (1994).
21. Spector, S. A., Gardiner, P. E., Zernicke, R. Z., Roy, R. R. & Edgerton, V. R. Muscle architecture and force-velocity characteristics of cat soleus and medial gastrocnemius: Implications for motor control. *J. Neurophysiol.* **44**, 951–960 (1980).
22. Cohen, A. H., Rossignol, S. & Grillner, S. *Neural Control of Rhythmic Movements in Vertebrates* 1–500 (Wiley, New York, 1988).
23. Jenkins, F. A. Jr The evolution of the avian shoulder joint. *Am. J. Sci.* **293A**, 253–367 (1993).
24. Sy, M. Funktionell-anatomische Untersuchungen am Vogelflügel. *J. Ornithol.* **84**, 253–267 (1936).
25. Hazelhurst, G. A. & Rayner, J. M. V. An unusual flight mechanism in the pterosauria. *Paleontology* **35**, 927–941 (1992).
26. Wellnhofer, P. in *Archaeopteryx* 1–30 *Ein neues Exemplar von Archaeopteryx* (Freunde des Jura-Museums Eichstätt, 1988).
27. Wellnhofer, P. in *Archaeopteryx* 1–47 *Das siebte Exemplar von Archaeopteryx aus den Solnhofener Schichten* (Freunde des Jura-Museums Eichstätt, 1993).
28. Sanz, J. L., Chiappe, L. M. & Buscalioni, A. D. The osteology of *Concornis lacustris* (Aves: Enantiornithes) from the lower Cretaceous of Spain and a reexamination of its phylogenetic relationships. *Novitates* **3133**, 1–23 (1995).
29. Kurochkin, E. N. A true carinate bird from Lower Cretaceous deposits in Mongolia and other evidence of Early Cretaceous birds in Asia. *Cretaceous Res.* **6**, 271–278 (1985).
30. Sanz, J. L. & Bonaparte, J. F. A new order of birds (class Aves) from the Lower Cretaceous (Spain). *Nat. Hist. Mus. L.A. County Sci. Ser.* **36**, 39–49 (1992).

Acknowledgements. We thank L. Chiappe, K. Earls, J. Gray-Chickering, C. Kovacs, F. A. Jenkins Jr, D. Ritter, J. Ostrom and T. A. McMahon for critically reviewing the manuscript and for their encouragement; M. Morimoto and A. Valore for technical assistance; and L. L. Mesozoely and K. Brown-Wing for Figs 1 and 3, and Fig. 2, respectively. This work was supported by a grant from the NSF.

Correspondence and requests for materials should be addressed to S.O.P. (e-mail: samuel_poore@brown.edu).

Doubling of world population unlikely

Wolfgang Lutz*, Warren Sanderson† & Sergei Scherbov‡

* International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria

† Department of Economics, State University of New York at Stony Brook, Stony Brook, New York 11794, USA

‡ Population Research Centre, Faculty of Spatial Sciences, University of Groningen, PO Box 800, NL-9700 AV Groningen, The Netherlands

Most national and international agencies producing population projections avoid addressing explicitly the issue of uncertainty. Typically, they provide either a single projection or a set of low, medium and high variants^{1,2}, and only very rarely do they give these projections a probabilistic interpretation. Probabilistic population projections have been developed for specific industrialized countries, mostly the United States, and are based largely on time-series analysis³. On a global level, time-series analysis is not applicable because there is a lack of appropriate data, and for conceptual reasons such as the structural discontinuity caused by the demographic transition⁴⁻⁶. Here we report on a new probabilistic approach that makes use of expert opinion on trends in fertility, mortality and migration, and on the 90 per cent uncertainty range of those trends in different parts of the world. We have used simulation techniques to derive probability distributions of population sizes and age structures for 13 regions of the world up to the year 2100. Among other things, we find that there is a probability of two-thirds that the world's population will not double in the twenty-first century.

The probabilistic projections are based on distributions for fertility, mortality and migration in all regions, defined in terms of high or low values assumed to cover 90 per cent of all possible future outcomes. For today's high-fertility countries they are based on an assessment of their current standing in the process of demographic transition towards low fertility⁷, together with information about reproductive intentions⁸. These data show that even in sub-Saharan Africa the fertility transition has started, and that it is well advanced in most other developing regions. The high and low assumptions for the years 2030–2035 are total fertility rates (TFR, the number of children per woman) of 4.0 and 2.0 in Africa, central Asia and the Middle East, 3.0 and 1.7 in southern Asia, Pacific Asia and Latin America and 3.0 and 1.5 in Central East Asia (mostly China).

For today's industrialized countries the assumptions are based on a broad survey of possible future societal changes⁹. The United

Nations and other institutions have assumed that fertility will eventually recover to replacement level (TFR slightly above 2.0), but there is little support for this view^{10,11}. Accordingly, TFR values of 2.1 and 1.3 for Europe and the Pacific members of the Organization for Economic Cooperation and Development (OECD countries) and 2.3 and 1.4 for North America have been assumed for 2030–2035.

Assumptions for mortality were set in terms of increase in life expectancy at birth per decade. Contrary to earlier beliefs, there now is a considerable degree of uncertainty about the future course of mortality. In the industrialized countries this stems from the scientific dispute of whether we are already close to a biologically determined limit to life expectancy^{12,13}. Accordingly, increases of 3.0 and 1.0 years have been assumed as the 90 per cent range. In the developing countries the uncertainty is more associated with future trends in AIDS¹⁴ and other infectious diseases and the development of health services¹⁵. For certain regions possible problems with food supply have also been considered^{16,17}. Consequently in such cases the assumed range of mortality improvement is wide, for example +4.0 to –2.0 years per decade in sub-Saharan Africa.

Migration is most difficult to handle because of unreliable data and high volatility¹⁸. For this study a matrix of constant annual interregional migration flows was assumed with the 90 per cent ranges covering two million to zero migration gains in North America, and 1 million to zero in Western Europe.

The projections for 13 regions (see Table 1) show that population growth will probably be most rapid in the middle East, sub-Saharan Africa and North Africa, with a tripling of the population by 2050 and a quadrupling by 2100 likely. Despite this rapid growth, there will also be significant increases in the proportion above 60 years of age. In contrast, in Eastern Europe and the European part of the former Soviet Union, population will probably decrease over the coming decades. By 2050 the Pacific OECD countries and Western Europe are likely to experience little, if any, change in population size. This stagnation or shrinkage in population size in Europe and the Pacific OECD countries will be associated with significant ageing of the population, with the proportion above 60 likely to double from its current values. Even proportions well above 40 per cent are within the 95 per cent confidence interval. These could bring serious consequences for social security systems.

In North America, a younger age distribution, a larger inflow of migrants, and slightly higher fertility than in Europe is likely to result in a roughly 25 per cent increase in population by 2050. Population ageing will occur, but will not be as dramatic as in Europe. Latin America is likely to have a doubling of its population and an increase in its proportion above age 60 to about 20 per cent. By 2050, Central East Asia (mostly China) is likely to grow by 37 per cent and experience an increase in the proportion over 60 from 9 per cent to 25 per cent. Southern Asia (essentially the Indian

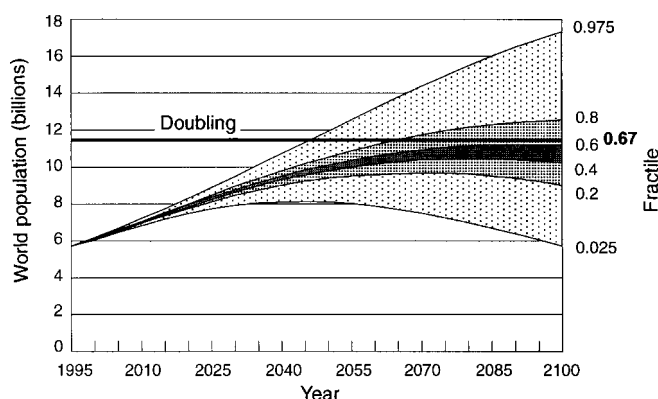


Figure 1 Fractiles of the probability distribution of the future size of the world population.

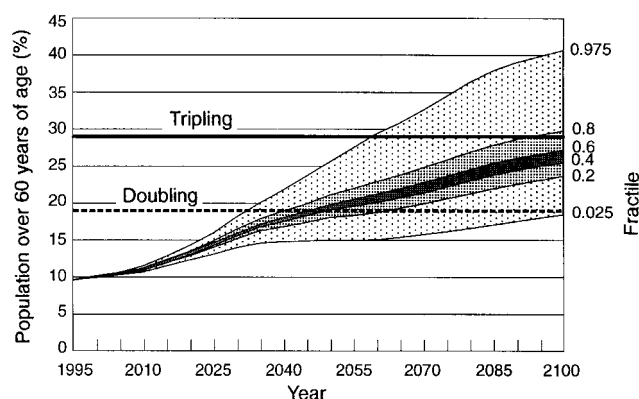


Figure 2 Fractiles of the probability distribution of the proportion of the world population above 60 years of age.

Table 1 Result of probabilistic projections for 13 world regions

Region	Total population (millions)				Population above age 60 (in %)			
	1995	2050			1995	2050		
		Median	2.5%*	97.5%*		Median	2.5%*	97.5%*
Africa								
North Africa	162	439	309	583	5.9	13.3	9.4	19.2
Sub-Saharan Africa	558	1,605	1,085	2,316	4.7	9.2	6.9	12.8
East Asia								
Central East Asia	1,362	1,865	1,351	2,574	9.2	24.9	17.8	34.1
Pacific Asia	447	796	579	1,047	6.8	19.4	14.4	26.5
Pacific OECD countries	147	146	117	182	19.4	39.5	31.5	48.7
West Asia								
Central Asia	54	137	88	206	7.8	15.4	10.2	24.0
Middle East	151	515	380	692	5.4	12.5	9.1	17.3
Southern Asia	1,240	2,368	1,833	2,970	6.7	16.6	13.4	20.8
Europe								
Eastern Europe	122	110	86	141	16.7	34.0	26.7	43.4
Former Soviet Union (European part)	238	188	144	241	16.9	34.1	26.3	44.5
Western Europe	447	471	370	584	18.6	35.0	27.5	43.9
Latin America	477	925	707	1,177	7.6	20.4	15.8	26.4
North America	297	403	303	534	16.4	30.2	24.0	38.6

Fertility, mortality and migration are assumed to be independent.

* Columns labelled 2.5% and 97.5% provide data on the lower and upper bounds, respectively, of the 95 per cent confidence interval.

subcontinent), which still has relatively high fertility and a young population, will probably double its population by 2050 and will be the world's most populous region.

The global results for population size are presented in Fig. 1; those for the proportion above 60 years of age are shown in Fig. 2. The median path of world population growth will increase from 5.8 billion today to 7.9 billion in 2020 and 10.0 billion in 2050. It will reach a peak around 2070–2080, and then begin a slow decline. In 2020 the range of uncertainty of this projection will be rather narrow (with the 95 per cent confidence interval between 7.5 billion and 8.3 billion) because many of the people who will be alive at that date have already been born. By 2050, the 95 per cent confidence interval will widen to between 8.1 billion and 11.9 billion, but the most likely 60 per cent (medium and dark shaded area in Fig. 1) still covers a range of less than one billion. After 2050 the size of the 95 per cent confidence interval will increase substantially, with the 60 per cent confidence interval showing a much smaller rise. The proportion of the world's population above age 60 is likely to increase from 9.5 per cent today to 20 per cent in 2050 to 27 per cent by 2100. A strong increase in the proportion of elderly people is virtually certain, with the low end of the 95 per cent confidence interval showing almost a doubling of today's level.

These probabilistic projections lead us to believe that the focus of public, political and scientific concern will continue to shift from global population growth to population ageing. □

Methods

The probabilistic population projections are based on the multistate cohort-component model of population projections, which applies assumed age-specific fertility, mortality and migration rates to the age and sex distribution of the starting population along cohort lines^{19,20}. A group of demographers have analysed trends in fertility, mortality and migration in different parts of the world²¹. Their discussions produced a consensus about ranges in 2030–2035 that they thought would cover 90 per cent of all future paths of TFR, life expectancy at birth, and the interregional migration matrix²². Because the resulting distributions of assumed values turned out to be symmetric, normal distributions were fitted to those ranges. For each of the three variables, a single draw from a standard normal distribution determined its relative position within its range of future values at selected dates. The values at intermediate dates were determined by piece-wise linear interpolation. This method has been labelled a random scenario approach to population projection²³.

Experiments with less autocorrelated paths for each variable produced very similar means and medians and minor differences in variances²⁴.

Beyond 2030–2035, fertility was assumed to reach an average level of between 1.7 and 2.1 children per woman by 2080, with the specific value depending on population density in 2030 (the higher the density, the lower the fertility). The 90 per cent range around that value was set at 1.0 children. The range for life-expectancy increases after 2030–2035 was set to 0–2.0 years per decade. Smooth transitions of assumed future life expectancies at birth into age-specific mortality rates were performed by transforming baseline mortality patterns with the help of relational models²⁵. Because separate baseline data were needed for males and females in each region, 26 patterns were used. Age-specific fertility rates were derived from the total fertility rates by using a fixed relative age profile of fertility. We used a fixed relative age profile for migrants to determine age-specific migration rates. The projections were performed in five-year steps on populations in five-year age groups.

An unusual feature of this global projection is the explicit consideration of possible correlations in fertility and mortality between and within regions. Four sets of 1,000 simulations were produced, resulting from a cross-classification of perfectly correlated/uncorrelated fertility and mortality trends across regions and within regions. The regional results presented in Table 1 are from the set with uncorrelated trends. The global results are based on the merged distribution of all four sets of simulations, which make up 4,000 projections in total.

Received 30 December 1996; accepted 17 April 1997.

1. United Nations *World Population Prospects: The 1994 Revision* (United Nations, New York, 1995).
2. United Nations *Long-Range World Population Projections: Two Centuries of Population Growth 1950–2150* (United Nations, New York, 1992).
3. Lee, R. D. Stochastic demographic forecasting. *Int. J. Forecast.* **8**, 315–327 (1992).
4. Nordhaus, W. *Managing the Global Commons* (MIT Press, Cambridge, MA, 1994).
5. Alho, J. Scenarios, uncertainty, and conditional forecasts of the world population. *J. R. Statist. Soc. A* **160**, Part 1, 71–85 (1997).
6. Lutz, W., Goldstein, J. R. & Prinz, C. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 14–44 (Earthscan, London, 1996).
7. Cleland, J. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 47–72 (Earthscan, London, 1996).
8. Westoff, C. F. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 73–87 (Earthscan, London, 1996).
9. Lutz, W. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 253–277 (Earthscan, London, 1996).
10. Keyfitz, N. in *Future Demographic Trends in Europe and North America. What Can We Assume Today?* (ed. Lutz, W.) 235–246 (Academic, London, 1991).
11. Westoff, C. F. in *Future Demographic Trends in Europe and North America. What Can We Assume Today?* (ed. Lutz, W.) 227–233 (Academic, London, 1991).
12. Vaupel, J. W. & Lundström, H. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 278–295 (Earthscan, London, 1996).
13. Manton, K. G. in *Future Demographic Trends in Europe and North America. What Can We Assume Today?* (ed. Lutz, W.) 97–115 (Academic, London, 1991).

14. Bongaarts, J. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 170–195 (Earthscan, London, 1996).
15. Garenne, M. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 149–169 (Earthscan, London, 1996).
16. Heilig, G. K. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 196–249 (Earthscan, London, 1996).
17. Cohen, J. E. *How Many People Can the Earth Support?* (Norton, New York, 1995).
18. Zlotnik, H. in *The Future Population of the World. What Can We Assume Today?* (ed. Lutz, W.) 299–335 (Earthscan, London, 1996).
19. Keyfitz, N. *Applied Mathematical Demography* (John Wiley, New York, 1977).
20. Rogers, A. *Introduction to Multiregional Mathematical Demography* (John Wiley, New York, 1975).
21. Lutz, W. (ed.) *The Future Population of the World. What Can We Assume Today?* Revised Edition. (Earthscan, London, 1996).
22. Lutz, W. *Scenario Analysis in Population Projection* (Working Paper WP-95-57, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1995).
23. Lee, R. D. Probabilistic approaches to population forecasting, in *Rethinking Population Projections* (eds Lutz, W. & Vaupel, J. W.) (International Institute for Applied Systems Analysis, Laxenburg, Austria) (submitted 1996).
24. Lutz, W. & Scherbov, S. *Sensitivity Analysis of Expert-Based Probabilistic Population Projections in the Case of Austria* (Working Paper, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1997).
25. Brass, W. in *Biological Aspects of Demography* (ed. Brass, W.) 69–110 (Taylor and Francis, London, 1971).

Acknowledgements. This work has been conducted at IIASA (International Institute for Applied Systems Analysis).

Correspondence and requests for materials should be addressed to W.L. (e-mail: lutz@iiasa.ac.at).

Attentional requirements in a 'preattentive' feature search task

Julian S. Joseph⁺, Marvin M. Chun[†] & Ken Nakayama[‡]

^{*} Department of Psychology, University of Nevada, Reno, Nevada 89557, USA

[†] Department of Psychology, Yale University, New Haven, Connecticut 06520, USA

[‡] Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA

It is commonly assumed that certain features are so elementary to the visual system that they require no attentional resources to be perceived. Such 'preattentive' features are traditionally identified by visual search performance^{1–3}, in which the reaction time for detecting a feature difference against a set of distractor items does not increase with the number of distractors. This suggests an unlimited capacity for the perception of such features. We provide evidence to the contrary, demonstrating that detection of differences in a simple feature such as orientation is severely impaired by additionally imposing an attentionally demanding rapid serial visual presentation task involving letter identification. The same visual stimuli exhibit non-increasing reaction time versus set-size functions. These results demonstrate that attention can be critical even for the detection of so-called 'preattentive' features.

One basic tenet of modern vision research is that certain attributes of visual stimuli can be processed and detected in parallel across the visual field^{4–6}. Visual attributes such as orientation^{1,5,6}, colour, or size differences³ have been put forth as 'preattentively' perceived stimulus properties, a concept introduced by Neisser⁴. Perhaps owing to the emphasis on a dichotomy between 'preattentive' and 'attentive' processing, it is commonly assumed that attentional resources are not necessary for the perception of such image properties. This dichotomy stems from a long history of research with the visual search paradigm in which the time to detect a target is measured as a function of the number of display items (reviewed in ref. 3). Stimulus attributes that require focal attention to be perceived exhibit positive slopes: reaction times increase with increasing display size. In contrast, some search tasks show reaction times that remain flat or even decrease slightly as the number of

items increases. Stimulus attributes leading to this behaviour, such as stimulus orientation, are thought to be processed in parallel across space with unlimited capacity; hence they are called 'preattentive' features, and are sometimes thought to be perceived without the use of attention. In the case of stimulus orientation, this fits well with the orientation selectivity of V1 neurons^{7,8}, which could conceivably permit the perception of orientation differences regardless of the attentional state.

If the perception of primitive features enjoys a special status in the visual processing stream, avoiding any bottleneck of limited resources, then performance in detection of a feature difference should be unaffected when attention is diverted elsewhere. We investigated the role of attention in the perception of 'preattentive' orientation features with a dual-task procedure as depicted in Fig. 1. We used a competing task, that of reporting the identity of a single white letter appearing in a rapidly changing stream of otherwise black letters at fixation. This rapid serial visual presentation (RSVP) is very demanding when presented at 12 letters s⁻¹ and has been shown to effectively consume attentional resources for periods up to half a second⁹. A search array of oriented Gabor patches was presented for 150 ms, immediately followed by 150 ms high-contrast white-noise masks covering their locations. The lag between the onset of the white target letter in the RSVP stream and the onset of the orientation array was randomly varied to examine the temporal extent of interference, if any. In the single-task condition, subjects were instructed to ignore the letters and report only whether an orientation 'oddball', a uniquely oriented item, had been present. In the dual-task condition, subjects were instructed to report both the white letter and whether an orientation oddball was present.

Severe impairments in performance in detecting orientation oddballs resulted when the attentionally demanding RSVP letter identification was additionally imposed (Fig. 2). In the condition of performing only the single task of orientation oddball detection, subjects performed well, averaging 94% correct. However, when performing the dual task of letter identification and orientation oddball detection, oddball detection accuracy was only 60 ± 5% for simultaneous letter and orientation onset (lag 0). Note that the chance level of performance is 50% in this task. Significant degradation in performance persists for several hundred milliseconds after the target letter's appearance, as a result of the attentional demands for processing the target letter^{9–11}. For the longest lag of 667 ms, dual-task performance recovered to the single-task level; thus the impairment reflects the temporal dynamics of attentional load rather than just a generic difficulty in encoding and retaining two independent responses. The effects of condition and lag, and the interaction between these variables were all significant ($P < 0.01$).

One might speculate that we observed attentional effects in the detection performance because these stimuli are unusual in some way and do not qualify as 'preattentive'. Performance was quite high when only the orientation oddball task had to be performed, but do these stimuli exhibit the standard experimental signature of so-called preattentive perception, specifically reaction times that do not increase with the number of items? There was no reason to expect otherwise, because many studies^{1,3} have found this for orientation differences that are easily detected. Our stimulus display, however, was slightly different from those used in the usual visual search task in that our search array was on for only a short fixed duration and was masked, whereas it is more customarily presented without a mask and for a longer duration up to the time of the observer's response. To remove any residual doubts, we did a visual search reaction time experiment on visual stimuli that were precisely the same as those used in the first experiment, the only difference being that the number of oriented Gabor items was varied from trial to trial. Subjects were instructed to ignore the letters and respond correctly on the presence of a uniquely oriented item as rapidly as possible. The letter stream was presented as well, although

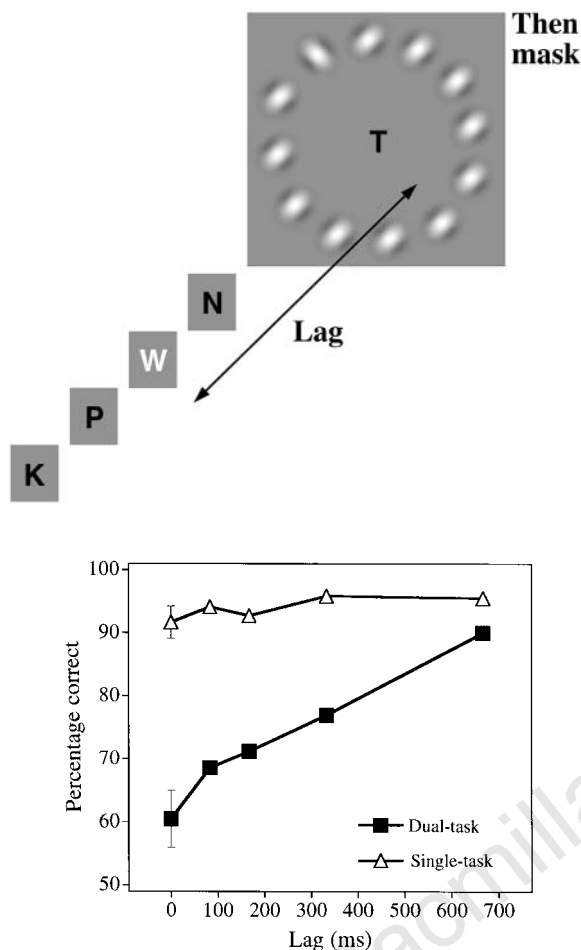


Figure 2 Percentage correct in the orientation oddball detection for the single-task (oddball detection only) and dual-task (letter identification and oddball detection) conditions as a function of the lag between the onsets of the target letter and the orientation array. Plots here and in subsequent figures represent the averages across subjects. Representative error bars indicate the s.e.m. across subjects. In the dual-task condition, the percentage correct in the oddball task was tabulated out of those trials in which the letter was identified correctly⁹ (letter accuracy was 83%; chance was 3.8%).

it was irrelevant to this task. Hence, on trials with set-size 12, the stimuli were physically identical in every respect to those used in the previous experiment.

The results (Fig. 3) show that the reaction time for detecting orientation oddballs does not increase with the number of distractor items. In fact, there is a slight decrease, which is to be expected from the higher density at the larger set-sizes. The perception of the same stimuli that displayed attentional requirements in the first experiment also exhibited the defining characteristic of a so-called 'preattentive' feature.

The circular arrays of oriented items discussed so far are the simplest way of studying the attentional requirements of detecting a feature difference, with each target and distractor having the same eccentricity and hence comparable visibility. One might imagine, however, that perception of an orientation difference within a two-dimensional array of items could conceivably be part of a qualitatively distinct class of visual tasks with no attentional requirements whatsoever. Such arrays were tested by Braun and Sagi^{12,13} who found no diminution of performance in highly practised subjects using a task of discriminating between the letters L and T at

Figure 1 Schematic of visual stimuli. The Gabor items were oriented at either $+45^\circ$ or -45° with respect to vertical. Half the trials contained one uniquely oriented item (an oddball) and half contained no oddball (all items identical). Subjects responded whether an oddball was present. An RSVP stream of letters (36 arcmin tall) was concurrently presented at fixation. In some blocks of trials, subjects had to report the white letter in addition to responding whether an oddball was present in the circular array (dual-task condition). In alternating blocks, subjects ignored the letters and responded only as to the presence of an orientation oddball (single-task condition). Responses were not speeded; subjects were instructed not to make their keypress responses until after the display sequence was completed.

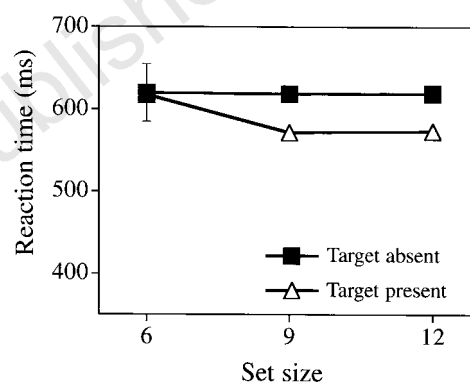


Figure 3 To check that the standard experimental signature of a preattentively perceived attribute was exhibited by our stimuli, we measured the detection reaction time as a function of the number of items in the display. The same stimuli were used as in the previous experiment, including the presentation of the letters, except that the number of oriented Gabor items was varied. Eight naive subjects each performed 10 blocks of 60 trials each, with oddball presence, distractor orientation, lag and set-size counterbalanced within each block. The reaction time for orientation oddball detection did not increase with the number of items in the search, which is characteristic of so-called 'preattentive' features. The error rate showed no set-size dependence and averaged 7%.

fixation¹³. To explore this issue with the RSVP letter identification task, we examined the detection of orientation differences in two-dimensional arrays of 36 oriented Gabor items (Fig. 4a) essentially identical to those used by Braun and Sagi¹², but with increased contrast (53% instead of 36%) and with longer Gabor duration (33 rather than 20 ms). A Gabor-mask stimulus onset asynchrony (SOA) of 200 ms was used. In all other respects the procedure was the same as in the first experiment. Profound attentional impairment was again observed (Fig. 4b) with the same qualitative pattern. The greatest impairment was at the shortest lag times after the target letter. We repeated this experiment with six different naive subjects for an SOA of 133 ms and stimulus duration 100 ms, obtaining the same effects (Fig. 4b). We also replicated Braun and Sagi's null result with the L/T discrimination task at a variety of orientation stimulus SOAs after subjects had received extensive practice in each single task and the dual task, although there was an effect in the initial blocks that quickly decreased as the tasks were learned. These observations suggest that extensive practice can greatly reduce the observable attentional effects, although the use of different attentional tasks, RSVP letter identification compared with L/T^{13,14}

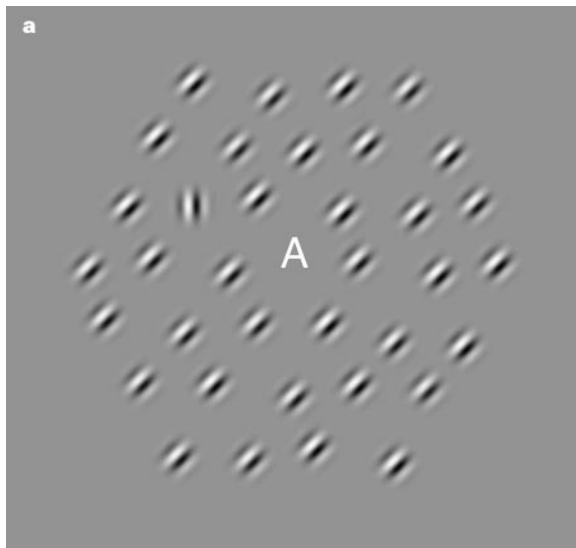
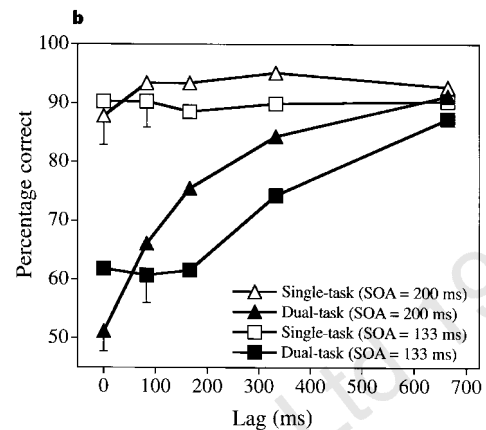


Figure 4 a, Two-dimensional array of oriented Gabor items used to demonstrate that similar performance impairments occur in the perception of this type of orientation difference. An orientation oddball, when present, was equally likely to appear in any location of the second concentric ring in the hexagonal array. The 33 ms array was followed by a 167 ms blank interval and a 100 ms array of masking elements consisting of superimposed high-contrast Gabor items of different orientations¹². Six naive subjects were tested in both single-task and dual-task



conditions, as in the first experiment. The same stimuli were tested on another six subjects with an SOA of 133 ms and a stimulus duration of 100 ms. **b**, In the two-dimensional arrays, detection of orientation differences was impaired by the attentionally demanding letter identification task (letter accuracy 76% and 75% respectively for the 200 and 133 ms SOA experiments). The effects of condition and lag, as well as the interaction between them, were all significant in each experiment ($P < 0.001$).

discrimination, might also have been partly responsible for the different results obtained in the present work.

In sum, our results show that visual feature search tasks, which have been deemed independent of attentional resources, are impaired by a sufficiently demanding attentional load. We demonstrated that both flat reaction time vs set-size functions and rapid discrimination performance are dissociated from unlimited-capacity detection of such 'preattentive' features. These findings seem to rule out a conceivable architecture for the visual system^{4-6,12,13} in which all feature differences are processed along a pathway that has a direct route to awareness, without having to pass through an attentional bottleneck. The results are consistent, however, with the theoretical notion¹⁵ that all tasks are contingent upon the availability of limited resources, while differing in their sensitivity to reductions in these resources. This view is supported by the results of Mack *et al.*^{16,17}, in which the presence of an otherwise salient object in the visual field frequently goes unnoticed when it is unexpected and irrelevant to the task at hand. In the present study, however, subjects are inherently constrained by limited attentional resources, even when actively interrogating the visual stimulus for the presence of a feature difference. These experiments argue against a direct route from preattentive processing to perceptual report. By providing a demonstration of 'preattentive' information that cannot be overtly perceived without attention, we challenge the current dichotomous view that assumes the existence of a separate privileged category of preattentive perception.

Methods

In the first experiment, six naive subjects, recruited for pay from the university community, each performed a 1-h session consisting of six blocks alternating between the single-task and dual-task conditions, with the order counter-balanced across subjects. Each block contained 80 trials, counterbalanced for the presence of an orientation oddball, distractor orientation and lag. Fourteen letters were presented after the target letter; between five and ten letters preceded it. No letter was repeated in a single trial. Luminances were 50 cd m^{-2} for the white letter, 7 cd m^{-2} for the black letters, and 25 cd m^{-2} for the background. Each letter in the stream was presented for 33 ms followed by a 50 ms blank gap. The Gabor functions were composed of a gaussian envelope

with 50% peak contrast and a standard deviation of 22 arcmin, with a cosine modulation of 110 arcmin wavelength. The Gabor items were regularly spaced around a circle at 5.3° eccentricity, with a random overall phase in the locations of the array. Stimuli were binocularly viewed from a distance of 57 cm. Each trial began with a small dot appearing at fixation for 500 ms, followed by a 500 ms blank interval and the beginning of the RSVP letter stream. Subjects were first shown frozen displays, and then given 20 trials of practice in each the single-task and dual-task conditions with feedback in the first 10 before the experiment began.

Received 25 November 1996; accepted 14 April 1997.

1. Treisman, A. M. Preattentive processing in vision. *Comput. Vis. Graph. Image Proc.* **31**, 156–177 (1985).
2. Treisman, A. M. & Sato, S. Conjunction search revisited. *J. Exp. Psychol.: Hum. Percept. Perf.* **16**, 459–478 (1990).
3. Wolfe, J. M. Guided search 2.0: A revised model of visual search. *Psychonom. Bull. Rev.* **1**, 202–238 (1994).
4. Neisser, U. *Cognitive Psychology* (Appleton-Century-Crofts, New York, 1967).
5. Julesz, B. Textons, The elements of texture perception, and their interactions. *Nature* **290**, 91–97 (1981).
6. Julesz, B. & Bergen, J. R. Textons, the fundamental elements in preattentive vision and perception of textures. *Bell Sys. Tech. J.* **62**, 1619–1645 (1983).
7. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)* **195**, 215–243 (1968).
8. Knierim, J. J. & Van Essen, D. C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophys.* **67**, 961–980 (1992).
9. Raymond, J., Shapiro, K. & Arnell, K. M. Temporary suppression of visual processing in an RSVP task: an additional blink? *J. Exp. Psychol.: Hum. Percept. Perf.* **18**, 849–860 (1992).
10. Duncan, J., Ward, R. & Shapiro, K. Direct measurement of attentional dwell time in human vision. *Nature* **369**, 313–315 (1994).
11. Chun, M. M. & Potter, M. C. A two-stage model for multiple target detection in rapid series visual presentation. *J. Exp. Psychol.: Hum. Percept. Perf.* **21**, 109–127 (1995).
12. Braun, J. & Sagi, D. Vision outside the focus of attention. *Percept. Psychophys.* **48**, 45–58 (1990).
13. Braun, J. & Sagi, D. Texture-based tasks are little affected by second tasks requiring peripheral or central attentive fixation. *Perception* **20**, 483–500 (1991).
14. Braun, J. Visual search among items of different salience: Removal of visual attention mimics a lesion in extrastriate area V4. *J. Neurosci.* **14**, 554–567 (1994).
15. Norman, D. A. & Bobrow, D. G. On data-limited and resource-limited processes. *Cogn. Psychol.* **7**, 44–64 (1975).
16. Mack, A., Tang, B., Tuma, R., Kahn, S. & Rock, I. Perceptual organization and attention. *Cogn. Psychol.* **24**, 475–501 (1992).
17. Rock, I., Linnett, C. M., Grant, P. & Mack, A. Perception without attention: results of a new method. *Cogn. Psychol.* **24**, 502–534 (1992).

Acknowledgements. This work was supported by the University of Nevada, Reno College of Arts and Science (J.S.J.), NIH (J.S.J., M.M.C.), and AFOSR (K.N.) and the McKnight Foundation (K.N.). We thank P. Cavanagh and J. M. Wolfe for helpful comments on the manuscript.

Correspondence and requests for materials should be addressed to J.S.J. (e-mail: jsj@unr.edu).

Restricted attentional capacity within but not between sensory modalities

John Duncan*, Sander Martens† & Robert Ward‡

* MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK

† Unit of Experimental and Theoretical Psychology, Faculty of Social Sciences, Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands

‡ School of Psychology, University of Wales, Bangor, Gwynedd LL57 2DG, UK

Restrictions to attentional capacity are revealed by the interference that commonly results when two sensory inputs must be identified at the same time¹. To investigate this phenomenon within and between modalities, we presented streams of visual and/or auditory inputs, containing occasional targets to be identified and recalled. For two visual or two auditory streams, identification of one target produced a sustained reduction in the ability to identify a second, the period of interference lasting for several hundred milliseconds. Subjectively, when attention was assigned to one target it was temporarily unavailable for another. In contrast, there was no such time-locked interference between targets in different modalities. The results suggest a modality-specific restriction to concurrent attention and awareness; visual attention to one simple target does not restrict concurrent auditory attention to another.

Rather generalized limitations on concurrent attention and awareness have often been proposed². According to such views, attention to any target stimulus or event should interfere with

concurrent attention to any other subsequent target, regardless of the targets' modalities. Very few studies, however, have compared interference within and between modalities, leaving the question unresolved. As one suggestive exception, Treisman and Davies³ found that searching for a target across two inputs in different modalities was more efficient than searching across inputs in the same modality. Here we extend this early study with exact measurements of the time-course of interference produced by one attended target on another, as a function of whether targets are presented within or between modalities. Interference between concurrent tasks can occur for many reasons⁴; measures of interference time-locked to a specific sensory input are useful in localizing the specific attentional demands of its processing⁴.

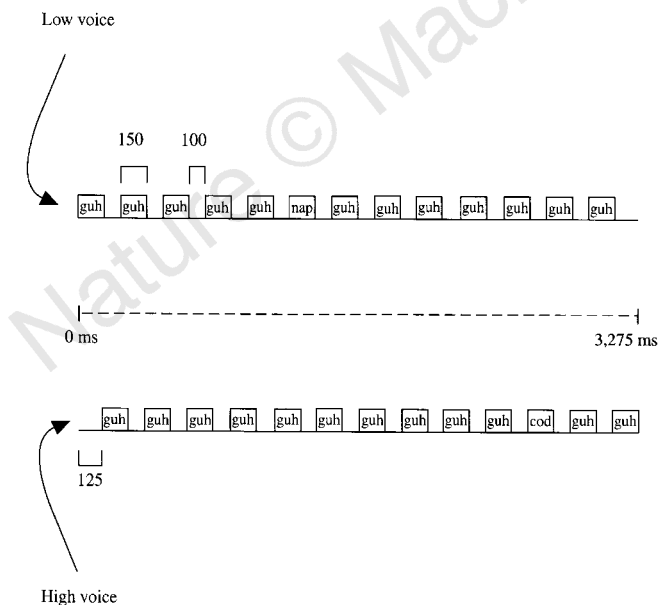


Figure 1 Example trial for single-modality auditory experiment. Two streams of speech were presented concurrently, one in a low and one in a high voice. Each stream consisted of identical repetitions of the syllable 'guh' (non-targets), with a single target word ('nap' or 'nab' for the low voice, 'cot' or 'cod' for the high voice) embedded within it. Syllables and words lasted for 150 ms each, and were separated by silent intervals of 100 ms. One stream, chosen at random on each trial, began 125 ms before the other, breaking synchrony of the two. The first stream to begin also contained the first target, presented after five non-targets. The second target was presented in the other stream, following delays of 125, 375, 625 or 1,375 (illustrated) ms measured from first to second target onset (stimulus onset asynchrony or SOA). After the second target, each stream was completed by two final non-targets.

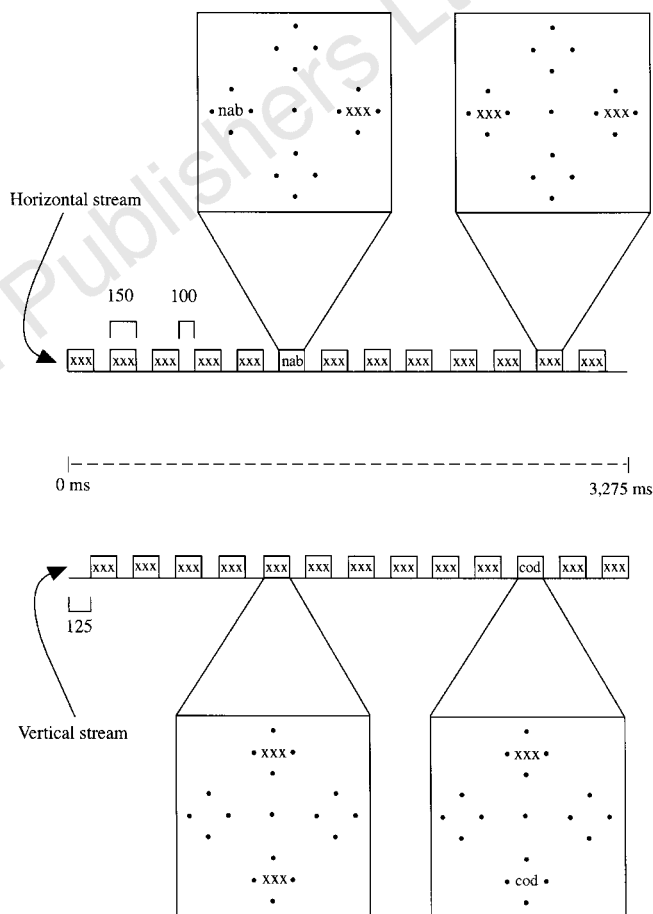


Figure 2 Example trial for single-modality visual experiment. Throughout each trial, the display contained a central fixation dot and four four-dot markers indicating stimulus positions. Overall display dimensions were ~4.3° square. As in the auditory experiment, there were two streams of input. Each stream was a series of 'frames', each presented for 150 ms and separated by a blank interval of 100 ms from the next. For the horizontal stream, each frame contained two letter strings, one in the left position and one in the right. In non-target frames, each string was a row of three xs, the whole string approximately 0.6° in length. In target frames, at random either the left or right non-target was replaced by the target word 'nap' or 'nab'. For the vertical stream, strings were presented in upper and lower positions, and the target was the word 'cot' or 'cod'. As indicated, frames for horizontal and vertical streams were partly overlapping in time. Thus, for part of the duration of any given frame, the other two display locations were empty (as illustrated), whereas for the remainder of the duration those locations also contained letter strings. Other details of timing were as in the auditory experiment, except that, to keep performance to an appropriate level, target words were presented for only 120 ms, and followed after this SOA by a 50 ms non-target (three xs) in the same location. SOA between the target and the next whole non-target frame, however, was preserved at the usual 250 ms.

Our methods followed those recently used to measure the time course or dwell time of visual attention⁵⁻⁷. Single-modality cases are illustrated in Figs 1 and 2. In the auditory case (Fig. 1), concurrent input streams were spoken by high and low voices. Each stream consisted of a string of non-targets (the spoken syllable 'guh'), with a single target word embedded somewhere within it. In the crucial divided-attention condition, the subject's task was to identify both targets; we measured the accuracy of identifying each target and, by varying their temporal separation, the time course of their mutual interference, or the dwell time of auditory attention. Focused-attention conditions, in which one stream of input was disregarded and only the target in the relevant stream was to be reported, were used as controls. The single-modality visual case (Fig. 2) was similar. Each stream of input was a series of briefly flashed letter strings presented one after the other. To ensure that central fixation was always maintained, the elements of each stream were not single strings but pairs of strings, either to left and right of fixation (horizontal stream) or above and below (vertical stream). Non-targets were rows of three xs, and targets again were words. Again the task in the crucial divided-attention condition was to identify both targets, one from the horizontal and one from the vertical stream.

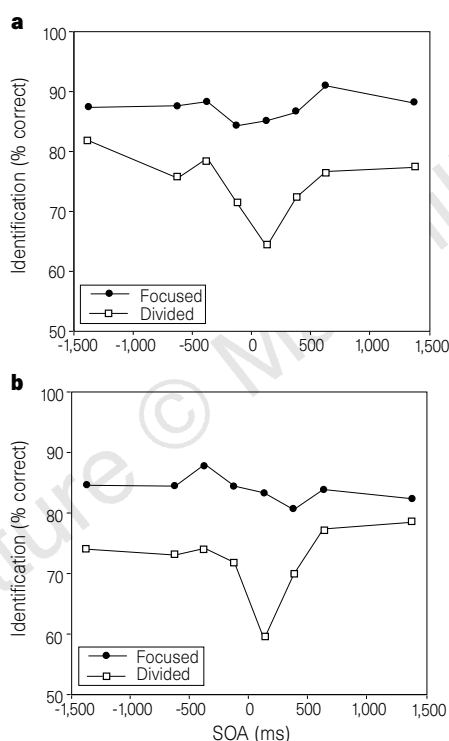


Figure 3 Mean accuracy (percentage correct word identification) in single-modality experiments. Negative SOAs refer to the first target in each trial; positive SOAs refer to the second target. For each experiment, the two streams (low and high voice for the auditory experiment, vertical and horizontal for the visual experiment) were scored independently; values shown are means across the two streams. **a**, Auditory experiment. Both the main effect of condition (focused compared with divided attention) and its interaction with SOA were significant (F -test, $P < 0.002$ in each case). In the divided-attention condition, accuracy at each SOA was compared with asymptote, defined as mean accuracy at SOAs of $\pm 1,375$ ms. Significant reductions occurred at SOAs of -125 , $+125$ and $+375$ ms, $P < 0.01$ or better. A final test showed a significant difference between focused and divided-attention conditions even at asymptote, $P < 0.01$. **b**, Visual experiment. Again, the main effect of condition and its interaction with SOA were both significant, $P < 0.001$ in each case. In the divided-attention condition, significant reductions compared with asymptote occurred at SOAs of $+125$ ($P < 0.001$) and (marginally) $+375$ ($P < 0.03$) ms. Again, focused and divided attention differed even at the asymptote, $P < 0.05$.

Figure 3 shows results from these single-modality experiments, comparing divided-attention conditions with focused-attention controls. Results in the two modalities were closely similar. Control conditions were more accurate overall, reflecting generalized dual-task decrements^{4,8,9} not time-locked to target onsets. Of primary interest is the time course of interference between one target and another in the divided-attention case. In either modality, accuracy was substantially reduced when a first target was followed within a few hundred milliseconds by a second target in the other stream. The results reflect the dwell time of attention on a first target and the subsequent release of attentional capacity for the next. Such interference was entirely absent from control conditions.

Figure 4 shows results from a mixed-modality experiment, in which one stream was auditory and the other was visual. Again there was a generalized dual-task decrement in the divided-attention condition. This time, however, there was absolutely no interference time-locked to target presentation. Attending to a target in one modality left concurrent identification of a second target in a different modality undisturbed.

We can consider the relation of these results to a variety of other phenomena in the dual-task literature. Previous findings of small overall differences between focused and divided attention, even

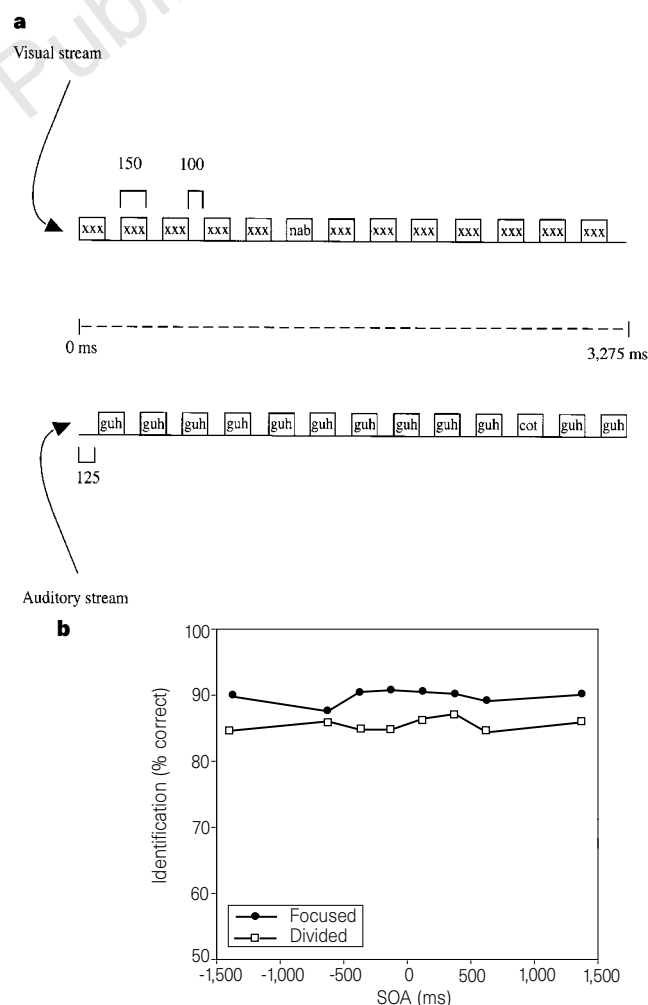


Figure 4 Mixed-modality experiment. **a**, Stimulus streams for a typical trial. The horizontal stream of the visual experiment was combined with the high voice stream of the auditory experiment. Timing details were as before. **b**, Mean accuracy. There was a significant main effect of condition (focused compared with divided attention), $P = 0.02$, but no main effect of SOA or condition-SOA interaction, each $F < 1$. Values are means for auditory and visual streams; except that the auditory task carried the bulk of the dual-task cost, auditory and visual data were similar.

when inputs are in different modalities^{3,10,11}, are consistent with our own generalized dual-task decrement (Fig. 4). As our results show, it is important that we measured interference time-locked to target presentation. Our experiments deal only with the case of independent visual and auditory inputs; a separate question concerns cross-modal integration when the same event gives rise to input in different modalities¹². Further sources of dual-task interference might also be important in more complex tasks, or when speeded responses are required⁴. Under the simple conditions of our experiments, however, attention to concurrent targets shows no cross-modal restriction. Neurophysiologically, the results suggest that a major source of attentional restriction must lie in modality-specific sensory systems. □

Methods

In each experiment, concurrent input streams were presented at a rate of 250 ms per item. A single target occurred in each stream. In focused-attention conditions, the task was to identify just the target in one specified stream. In divided-attention conditions, both targets were to be identified.

Streams for the single-modality auditory experiment are illustrated in Fig. 1. To create these sequences, single instances of each target and non-target were recorded, then cut to a length of 150 ms and combined into appropriate streams with Macromedia SoundEdit Pro. These streams were presented over headphones at a loudness of approximately 78 dB sound pressure level, with an Apple Macintosh IIci running Cedrus SuperLab software.

Streams for the single-modality visual experiment are illustrated in Fig. 2. Stimuli in this experiment were presented on an Apple Power Macintosh 8500 running Psyscope software¹³. Exact measurements of display timing revealed small variations around the values specified in the program; for example, when stimulus onset asynchrony (SOA) between the two targets was specified as 125 ms, measured values ranged from 119 to 148 ms, and similar levels of variability were seen in target durations and in SOA between one frame and the next within each stream. Such fluctuations were the same under divided- and focused-attention conditions and can accordingly be ignored.

Streams for the mixed-modality experiment are illustrated in Fig. 4. The experiment was run on an Apple Power Macintosh 8500 running Psyscope software, with temporal variability comparable to the visual experiment.

In all experiments, the subject initiated a trial by pressing the space bar of the computer keyboard. Stimulus streams began after a fixed delay of 250 ms; identification responses were typed in after the streams finished, using keys appropriately labelled for the targets 'cot', 'cod', 'nap' and 'nab'. Under focused-attention conditions, there was a single response identifying the attended target (two-alternative forced choice). Under divided-attention conditions, two responses were typed in, in either order. Subjects were strongly encouraged to take time over their responses, ensuring that typing errors were not made. For comparability, the fixation display (central dot and location markers, Fig. 2) was presented in all three experiments, and central fixation was required throughout the trial.

Each subject served in two focused-attention conditions, one for each stream, and the divided-attention condition. Each experiment began with three practice blocks of 32 trials, one per condition. In the auditory experiment there followed a first set of three experimental blocks (96 trials each), one per condition, followed by a second, similar set; other experiments had only one set of three experimental blocks (128 trials each), one per condition. The order of conditions was fixed for any one subject but counterbalanced across subjects. The experiment was conducted in a single session lasting ~1.5 h.

Single-modality auditory and visual experiments had 12 paid subjects each, aged 19–50 years. The mixed-modality experiment had 18 subjects, aged 19–49 years.

Received 24 January; accepted 15 April 1997.

1. Broadbent, D. E. *Perception and Communication* (Pergamon, London, 1958).
2. Posner, M. I. & Petersen, S. E. The attention system of the human brain. *Annu. Rev. Neurosci.* **13**, 25–42 (1990).
3. Treisman, A. M. & Davies, A. in *Attention and Performance IV* (ed. Kornblum, S.) 101–117 (Academic, London, 1973).
4. Pashler, H. Dual-task interference in simple tasks: data and theory. *Psychol. Bull.* **116**, 220–244 (1994).
5. Duncan, J., Ward, R. & Shapiro, K. Direct measurement of attentional dwell time in human vision. *Nature* **369**, 313–315 (1994).
6. Broadbent, D. E. & Broadbent, M. H. P. From detection to identification: response to multiple targets in rapid serial visual presentation. *Percept. Psychophys.* **42**, 105–113 (1987).

7. Raymond, J. E., Shapiro, K. L. & Arnell, K. M. Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perf.* **18**, 849–860 (1992).
8. Logan, G. D. Attention in character-classification tasks: evidence for the automaticity of component stages. *J. Exp. Psychol. Gen.* **107**, 32–63 (1978).
9. Bourke, P. A., Duncan, J. & Nimmo-Smith, I. A general factor involved in dual task performance decrement. *Q. J. Exp. Psychol.* **49A**, 525–545 (1996).
10. Lindsay, P. H., Taylor, M. M. & Forbes, S. M. Attention and multi-dimensional discrimination. *Percept. Psychophys.* **4**, 113–117 (1968).
11. Massaro, D. W. & Warner, D. S. Dividing attention between auditory and visual perception. *Percept. Psychophys.* **21**, 569–574 (1977).
12. Driver, J. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* **381**, 66–68 (1996).
13. Cohen, J. D., MacWhinney, B., Flatt, M. & Provost, J. Pyscope: a new graphic interactive environment for designing psychology experiments. *Behav. Res. Meth. Inst. Comp.* **25**, 257–271 (1993).

Acknowledgements. We are grateful to Christopher Robinson and Sally Cox for initial work on this project, and to Sophie Scott and Christian Lorenzi for assistance with stimulus generation and measurement.

Correspondence and requests for materials should be addressed to J.D. (e-mail: john.duncan@mrc-apu.cam.ac.uk).

The small GTP-binding protein Rab3A regulates a late step in synaptic vesicle fusion

Martin Geppert^{*}, Yukiko Goda[†], Charles F. Stevens[‡] & Thomas C. Südhof[‡]

^{*} Max Planck Institute for Experimental Medicine, Hermann-Rein-Strasse 3, 37075 Göttingen, Germany

[†] Molecular Neurobiology Laboratory and Howard Hughes Medical Institute, The Salk Institute, La Jolla, California 92037, USA

[‡] Department of Molecular Genetics and Howard Hughes Medical Institute, The University of Texas Southwestern Medical Center, Dallas, Texas 75235, USA

The Rab family of low-molecular-mass GTP-binding proteins are thought to guide membrane fusion between a transport vesicle and the target membrane, and to determine the specificity of docking^{1–3}. The docking and fusion of vesicles is, however, a complex multistep reaction, and the precise point at which Rab proteins act in these sequential processes is unknown. In brain, the Rab protein Rab3A is specific to synaptic vesicles, whose exocytosis can be monitored with submillisecond resolution by following synaptic transmission. We have now determined the precise point at which Rab3A acts in the sequence of synaptic vesicle docking and fusion by using electrophysiological analysis of neurotransmitter release in Rab3A-deficient mice. Unexpectedly, the size of the readily releasable pool of vesicles is normal, whereas Ca²⁺-triggered fusion is altered in the absence of Rab3A in that a more-than-usual number of exocytic events occur within a brief time after arrival of the nerve impulse.

We first investigated whether the *rab3A* deletion modifies the properties of the fundamental unit of transmitter release: the quantity of transmitter contained in each vesicle and the basal rate of spontaneous vesicle fusion. We used spontaneous miniature excitatory postsynaptic currents (mEPSCs), each one of which corresponds to exocytosis by an individual vesicle⁴ (the neurotransmitter quantum), to compare the amplitudes of individual quantal responses between wild-type and *rab3A* mutant synapses (Fig. 1a). Both the amplitude of mEPSCs and their basal spontaneous rate are unchanged at the *rab3A*-mutant synapses (Fig. 1b).

With rapid repeated use, the average quantity of neurotransmitter released by a synapse declines, a decline that probably represents the depletion of fusion-competent vesicles in the 'readily releasable pool' at the presynaptic terminal⁵. This depletion is faster in *rab3A*-mutant neurons, indicating some alteration in the efficiency of synaptic vesicle trafficking⁶. To determine whether the depletion is faster because the readily releasable pool is smaller than usual, we compared the pool size at synapses formed between cultured hippocampal neurons prepared from wild-type or *rab3A*-mutant mice.

The size of the readily releasable pool was estimated by applying hypertonic solution which induces vesicular release through a Ca^{2+} -independent mechanism and sharply increases the rate of occurrence of mEPSCs⁷. When a synapse is placed in a hyperosmotic environment (solution made hyperosmolar with 0.5 M sucrose), quanta are at first released rapidly and then, as the readily releasable pool is depleted, at progressively slower rates (Fig. 2a). The total number of quanta released per synapse during this burst of exocytosis defines the readily releasable pool, whereas the steady-state exocytic rate corresponds to the balance of release and recruitment of new vesicles to the readily releasable pool. The pool size of readily releasable vesicles, determined by integrating the release transient (Fig. 1 legend), is not appreciably altered in *rab3A*-mutant cells. The result from a synaptotagmin I mutant is included for comparison (Fig. 2c). In contrast to the reduction in evoked quantal release caused by the lack of fast Ca^{2+} -dependent transmitter release⁸, Ca^{2+} -independent release induced by hypertonic solution is intact in synaptotagmin I mutant cells (releasable pool size is 25–36 vesicles), confirming the specific role for functional synaptotagmin I in Ca^{2+} -dependent fusion processes.

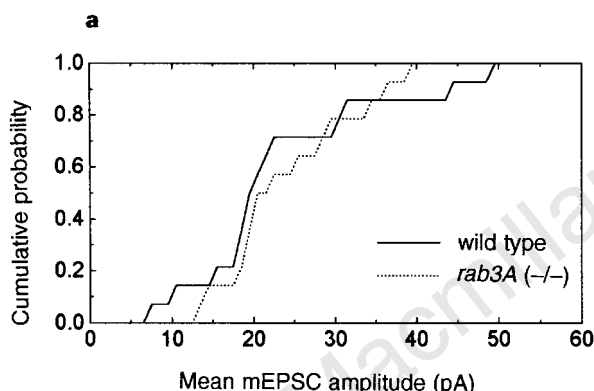
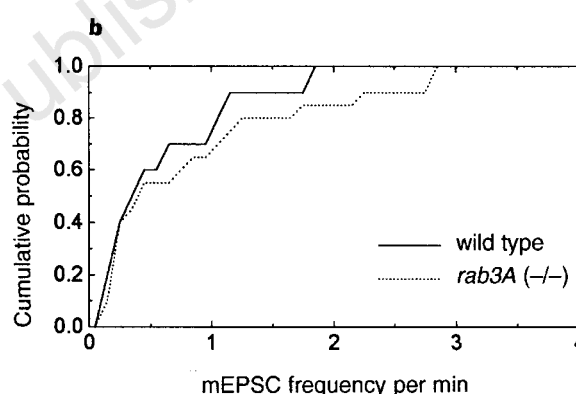


Figure 1 Spontaneous transmitter release is normal in the absence of Rab3A. Cumulative probability distributions of mean mEPSC amplitude (**a**) and frequency (**b**) from wild-type (straight line) and *rab3A* mutant (dotted line) neurons. Spontaneous mEPSCs were collected from hippocampal neurons grown in culture. The overall mean mEPSC amplitudes were 23 ± 3 pA ($\pm$ s.e.m.; $n = 14$) in wild-type cells, and 22 ± 2 pA ($\pm$ s.e.m.; $n = 14$) in mutant cells. The overall mean

Although the size of the release-ready vesicle pool in Rab3A-deficient mice is normal, a change in the kinetics of vesicle recruitment—such as vesicle docking and priming—could reflect slower refilling of the pool. We next examined such a possibility by estimating the refilling rate of the readily releasable pool in *rab3A*-mutant cultures with double-pulse application of hypertonic solution. The peak response to the second of a pair of hypertonic solution pulses applied at short interpulse intervals recovers with a time constant of ~ 7 s (ref. 7). This measure corresponds to the rate of replenishing releasable vesicles to a functionally defined pool that is shared by both nerve-impulse-evoked and hypertonic-solution-induced fusion processes⁹. As shown in Fig. 2d, the rate of recovery between successive applications of hypertonic solution is comparable for wild-type and *rab3A*-mutant neurons; *rab3A*-knockout mice are therefore not deficient in ensuring the continual supply of synaptic vesicles for fusion. Synaptotagmin I-mutant neurons also show an unaltered refilling rate of the readily releasable pool.

We have shown so far that the size and the refilling rate of the readily releasable vesicle pool are normal in *rab3A*-mutant mice, so



mEPSC frequencies were 0.62 ± 0.18 per min per synapse ($\pm$ s.e.m.; $n = 10$ neurons) in wild-type and 0.88 ± 0.20 per min per synapse ($\pm$ s.e.m.; $n = 20$ neurons) in mutant cells. The distributions between the wild type and Rab3A mutants of both mEPSC amplitude (**a**) and frequency (**b**) are not significantly different, as determined by Kolmogorov-Smirnov test ($P > 0.2$).

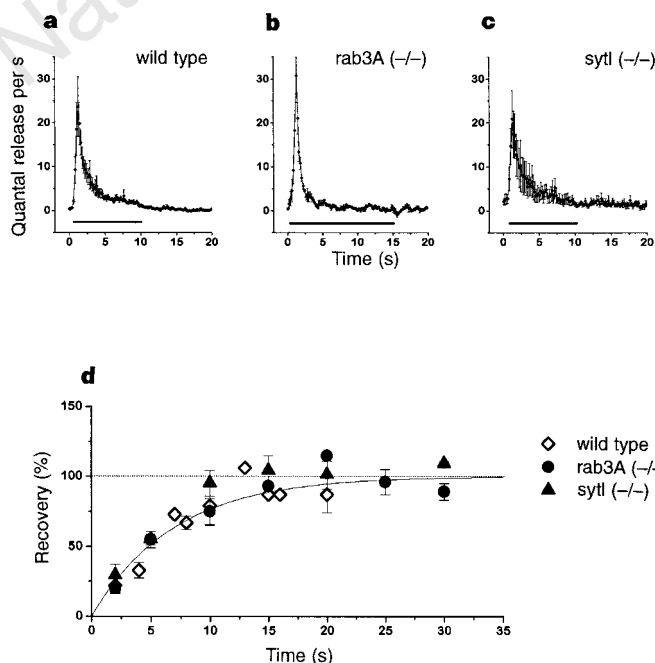


Figure 2 The size and refilling rate of the readily releasable pool are unchanged in *rab3A*-mutant synapses **a–c**, *rab3A*-mutant cells display a normal sized pool of release-ready vesicles. Average quantal release rate per synapse is shown as a function of time before and during local application of hypertonic solution (solid bar; external solution to which 0.5 M sucrose was added) for wild-type (**a**; $n = 6$), *rab3A*-mutant (**b**; $n = 5$), and synaptotagmin I (*sytl*)-mutant (**c**; $n = 3$) cells. Experiments were done in cultured hippocampal neurons. Bath solution contained 3 mM Ca^{2+} , 2 mM Mg^{2+} and 0.5 μM tetrodotoxin (TTX). Error bars represent s.e.m. There is no appreciable difference in the rate profile of quantal release induced by hypertonic solution among the three groups. The estimated releasable pool size was 26–35 vesicles in wild type, 28–31 vesicles in *rab3A* mutant, and 25–36 vesicles in synaptotagmin I mutant cells (see Methods). **d**, Recruitment of vesicles to the release-ready pool is not defective in *rab3A* mutant cells. Pairwise application of hypertonic solution was used to compare the recovery rates of the releasable pool of vesicles in cultured hippocampal neurons. Peak responses to the second hypertonic solution applied at various intervals are shown as a percentage of the peak response to first application. Points represented by wild-type cells (diamonds; $n = 11$), *rab3A*-mutant cells (circles; $n = 9$), and synaptotagmin-I-mutant cells (triangles; $n = 7$) have been fitted to an exponential recovery curve with a time constant of 6.9 s (solid line). Error bars indicate s.e.m. Wild-type, *rab3A*-mutant and synaptotagmin-I-mutant cultures show similar recovery rates of the releasable pool of vesicles.

the steps leading to the final Ca^{2+} -triggered exocytosis must be unaffected by the absence of *rab3A*. Moreover, the quantal properties are also unaffected by the mutation. The faster depletion with repeated stimuli in *rab3A* mutants⁶ could then be due to an increased efficiency of transmitter release⁵. We therefore tested whether synaptic vesicle fusion is affected in the mutant mice by comparing the quantity of transmitter released per synapse in wild-type and mutant neurons.

We monitored synaptic transmission from a known number of synapses in wild-type or *rab3A*-mutant hippocampal cultures to determine the average amount of transmitter release per synapse for every stimulus (Fig. 3a). A typical wild-type synapse releases, on average, about 0.5 quanta per stimulus under our recording conditions. In contrast, *rab3A*-mutant cells show a twofold increase in peak quantal release (one quantum per synapse per stimulation). Lack of Rab3A thus leads to an increase in the average number of vesicle fusions per synapse for every stimulus. Figure 3a also shows the average peak quantal release for synaptotagmin I mutant neurons for comparison. The greatly reduced efficiency of transmitter release in synaptotagmin I mutant cells, caused by disruption of the mechanisms supporting fast Ca^{2+} -dependent synaptic transmission, is reflected in the diminished peak quantal release per synapse relative to wild-type synapses⁸.

Although the average number of quanta released in response to nerve stimulation is increased, the size and refilling rate of the readily releasable pool, as well as the basal quantal release rate, remain normal in *rab3A*-mutant synapses. Contrary to the presumed role for Rab proteins in vesicle docking, Rab3A protein function therefore involves some step whose modification becomes apparent only at the time of evoked neurotransmitter release.

To determine whether Rab3A alters release properties in a time window immediately after stimulation, we tested the effect of Rab3A on a form of short-term plasticity that results from directly modulating Ca^{2+} -dependent exocytosis. Paired-pulse facilitation is an enhancement of the response to the second of a pair of stimuli given at short interpulse intervals (Fig. 3b, circles). This facilitation reflects an increase in release in response to the second stimulus caused by the previous exposure of the presynaptic terminal to a build-up of Ca^{2+} ions during the first pulse¹⁰. Facilitation is most prominent at shorter intervals, when the effects of Ca^{2+} accumula-

tion from the first stimulation have the strongest influence on transmitter release caused by the second pulse¹¹.

Facilitation at *rab3A*-mutant synapses is much larger than at wild-type synapses at short interpulse intervals (10–20 ms; Fig. 3b, triangles). The *rab3A*-mutant phenotype is most markedly expressed within 20 ms, starting with the onset of release to the first pulse, and is evident as increased release during this period. We have already shown that *rab3A*-mutant synapses release, on average, more quanta for every stimulus (Fig. 3a). The fact that *rab3A*-mutant synapses facilitate even more than expected when probed with a second closely spaced stimulus indicates that the mutant phenotype is enhanced by some aspect of the first release; otherwise, if mutant synapses simply release more quanta for every stimulus, then the ratio of the second response to the first response (that is, the paired-pulse facilitation ratio) should have remained unchanged.

We define transmitter-release probability, P , as the probability (reliability) that one or more quanta are released from a synapse in response to nerve stimulation. An increased quantal release in *rab3A*-mutant synapses could then reflect either an increase in P of releasing one or more quanta, and/or an increase in the number of quanta released when a release does occur (note that P remains unchanged in the latter case). We therefore compared the kinetics of inhibition of synaptic responses by MK-801 for wild-type and *rab3A*-mutant slices in two ways. First, we measured the probability of exocytosis for a given synapse population. Second, we repeated the experiment under conditions in which we could detect the difference in the amount of transmitter released for a synapse population.

MK-801 is effectively an irreversible open-channel blocker of the NMDA (*N*-methyl-D-aspartate) subtype of glutamate receptors¹². A progressive blocking of NMDA-receptor-mediated synaptic currents by MK-801 is used to estimate the transmitter-release probability P . For example, a higher value of P causes more synaptic vesicle fusions so that, on average, more transmitter molecules are released for each trial; under such conditions more synaptic NMDA receptors are opened for every stimulus, leading to a faster rate of inhibition of NMDA-receptor-mediated synaptic currents by MK-801 (refs 13, 14). Changes in the MK-801 blocking rate are evident with several-fold changes in release probability¹⁵. Consequently, if

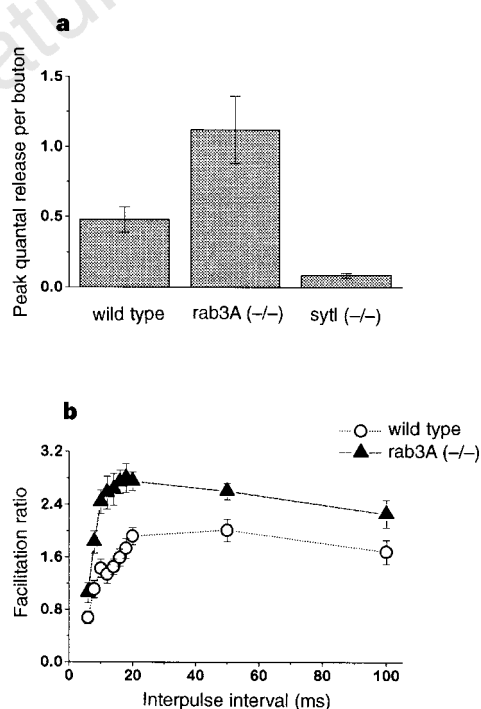


Figure 3 Mutation in *rab3A* involves alterations closely related to the final step of exocytosis. **a**, Peak quantal release is selectively enhanced in *rab3A*-mutant cultures. Average peak quantal release per synapse was determined by monitoring synaptic transmission at 0.1 Hz from a restricted number of synapses through local microperfusion⁸. Culture preparation was used for this experiment. Peak quantal release (peak current per quantal size) was 0.48 ± 0.09 for wild-type cell pairs ($n = 14$), 1.12 ± 0.24 for *rab3A*-mutant neuron pairs ($n = 14$), and 0.086 ± 0.016 for synaptotagmin-I mutant cells pairs ($n = 7$)⁸. Whereas the lack of functional Ca^{2+} sensor causes a severe reduction in peak quantal release, the absence of Rab3A protein increases the amount of transmitter released per stimulus. Error bars indicate s.e.m. **b**, Paired-pulse facilitation is enhanced in *rab3A* mutants. EPSCs to pairs of stimuli were monitored in wild-type (circles: $n = 15$ –22 slices from 3–6 mice) and *rab3A*-mutant (triangles: $n = 16$ –26 slices from 4–7 mice) hippocampal slices in extracellular recording solution that contained 2.5 mM Ca^{2+} and 1.3 mM Mg^{2+} . The facilitation ratio is the ratio of the second response to the first response, and is shown as a function of interpulse interval. Error bars give s.e.m. Paired-pulse response is greater for *rab3A* mutant neurons at all interpulse intervals tested, and is seen consistently under all conditions that alter initial release probability—that is, in four different extracellular Ca^{2+} and Mg^{2+} concentrations (5, 2.5, 1.25 mM Ca^{2+} at 1.3 mM Mg^{2+} , and 1 mM Ca^{2+} /2.4 mM Mg^{2+} ; data not shown). A previous study of *rab3A*-knockout mice included an analysis of paired-pulse facilitation; however, paired-pulse responses were examined for longer interpulse intervals only (≥ 50 ms; Fig. 2b, ref. 6). We focused our analysis on shorter interpulse intervals, primarily between 6 and 20 ms, when the effect of *rab3A* mutation on paired-pulse facilitation is most prominent.

the twofold increase in peak release in *rab3A* mutants (Fig. 3a) is due to a twofold increase in P , then we should detect a faster MK-801 blocking rate in the mutants under normal recording conditions. However, the rates of MK-801 blocking are the same in wild-type and *rab3A*-mutant slices (Fig. 4a). The increase in quantal release in the *rab3A* mutants is therefore not caused simply by a general increase in P across all synapses.

As peak quantal release increases but MK-801 blocking rate is unaffected, transmitter release in the mutants might be raised non-uniformly across synapses. Specifically, increased vesicle fusion must occur at synapses that already release at least one quantum (that is, those synapses associated with a higher value of P) because fusion of a single quantum under ordinary conditions is believed to saturate postsynaptic NMDA receptors^{16,17}. The sensitivity to MK-801 would therefore be unaffected only when additional quanta are released at synapses—that is, if *rab3A* mutation causes an increase in the number of quanta released per synapse instead of a higher number of synapses that release a quantum. To increase the detection limit of postsynaptic NMDA receptors to released transmitters, we did the MK-801 experiments in the presence of a high-affinity competitive antagonist of the NMDA receptor, D-2-amino-5-phosphonovaleate (AP5)¹⁸. Addition of 5 μ M AP5 reduces the effective agonist concentration and thus decreases the number of open NMDA receptors available for MK-801 binding. Consequently, the MK-801 blocking rate in the presence of AP5 is greatly slowed in wild-type slices (circles in Fig. 4a and b, circles). Under these conditions, the increased release of glutamate should reduce the potency of the competitive antagonist, leading to more open NMDA receptors and resulting in faster MK-801 blocking. In the presence of AP5, we find that the MK-801 block is indeed faster in *rab3A*-mutant slices than in wild-type slices (Fig. 4b), consistent with increased quantal release in *rab3A*-mutant synapses (Fig. 3a). Moreover, the fact that MK-801 blocking is faster in *rab3A* mutants only in the presence of AP5 is in accord with the saturation of postsynaptic NMDA receptors under the normal conditions used for monitoring transmitter release. The different MK-801 blocking rates in the absence (unaltered) and presence (faster) of AP5 in *rab3A*-knockout mice relative to wild-type mice, are evidence for an increased number of quantal releases per synapse and not a general increase in the number of synapses that release a quantum.

We have shown that the average quantal release per synapse is increased in the absence of Rab3A. Rab3A, therefore, is normally involved in regulating the number of quanta released at hippocampal synapses. A similar role for Rab3A in regulating fusion in chromaffin cells has been proposed^{19,20}. We have also determined when Rab3A functions during the different stages of synaptic vesicle exocytosis. Contrary to the popular view that Rab proteins are involved in vesicle docking, the effect of the *rab3A* mutation is specific to the time of nerve-impulse-evoked vesicle fusion; processes leading to Ca^{2+} -triggered exocytosis remained unchanged. The specificity of Rab3A function for nerve-evoked release is supported by the finding that spontaneous release is normal and also by a previous report that Rab3A dissociates from synaptic vesicles upon exocytosis only if release is triggered by Ca^{2+} influx²¹. Although Rab3A functions in regulating vesicle fusion at the time of the Ca^{2+} signal, Rab3A and/or its possible effectors may become incorporated or act on the fusion complex at an earlier stage, before the influx of Ca^{2+} ions.

How is quantal release increased in the absence of Rab3A? We believe that this increase is not caused by a general increase in release probability, but by more vesicle fusions when a release does occur because (1) occlusion or attenuation of paired-pulse facilitation, which is expected of processes that rely on increasing P (refs 22–24), was not observed in the mutants; rather, at all interpulse intervals tested, *rab3A*-mutant slices showed greater facilitation than wild-type slices (Fig. 3b), compatible with an unaltered value of P in the absence of Rab3A; (2) the fact that a difference in MK-801 blocking rates between wild-type and mutant slices was revealed only in the presence of a competitive antagonist of NMDA receptors indicates that quantal release is increased without an overall increase in P in the mutants. In addition, the MK-801 results argue for the saturation of NMDA receptors during synaptic transmission under normal conditions. Altogether, unaltered P at *rab3A* mutant synapses suggests that Rab3A regulates the efficiency of the fusion process after activation of the Ca^{2+} -sensitive step that governs the release probability.

We propose a mechanism by which Rab3A might regulate vesicle fusion. Transmitter release at hippocampal synapses is believed to be limited to a single quantum, even under conditions of maximal release probability^{25–28}. We suggest that Rab3A plays a role in the

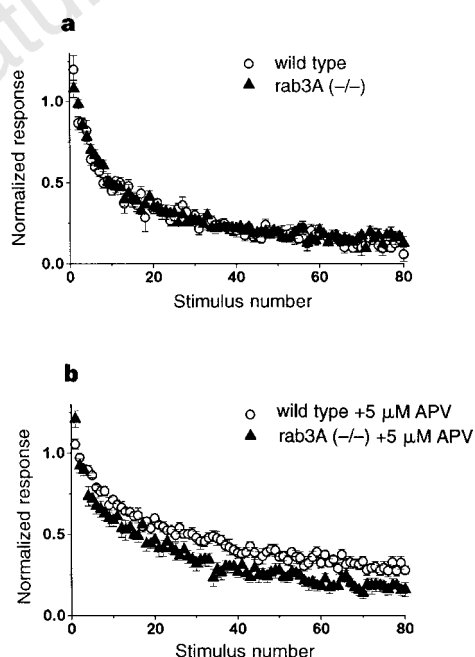


Figure 4 In the absence of Rab3A, the probability of transmitter release is not altered but the amount of transmitter released is increased. **a**, NMDA-receptor component of whole-cell synaptic currents was monitored in the presence of 5 μ M CNQX and 40 μ M MK-801 from wild-type (circles; $n = 6$ slices from two mice) and *rab3A*-mutant (triangles; $n = 7$ slices from two mice) hippocampal slices. The charge carried by the NMDA receptor currents was normalized to the intercept value obtained from linear regression of the first 11 responses, and is plotted as a function of stimulus number; synaptic responses were monitored every 10 s (ref. 15). Extracellular recording solution contained 2.5 mM Ca^{2+} and 1.3 mM Mg^{2+} . Error bars are s.e.m. The progressive blocking rate of NMDA-receptor-mediated synaptic currents by MK-801 is indistinguishable in wild-type and *rab3A* mutant cells, indicating that release probability is unaltered in the mutant animals. **b**, Comparison of the MK-801 blocking rate in the presence of 5 μ M APV reveals a faster rate in the mutant cells (triangles; $n = 10$ slices from five mice) relative to wild-type neurons (circles; $n = 10$ slices from three mice), consistent with increased quantal release at individual synapses in the absence of Rab3A.

mechanism that ordinarily limits release to a single quantum: suppose that such regulation works by raising the energy barrier for neighbouring vesicle fusion as a result of rapid mechanical deformation of the active zone, caused by insertion of additional membrane by the fused vesicle²⁹. Rab3A or its effectors could secure such rapid deformation of the active zone by facilitating the complete and efficient fusion of the synaptic vesicle membrane through destabilizing the metastable intermediate fusion complex. A role for temporal regulation of the stability of the docking and/or fusion complex has also been proposed for the endosome-specific Rab family member Rab5 (ref. 30). Then, in the absence of Rab3A, the precise steps are slowed that lead to complete fusion of the synaptic vesicle upon binding Ca^{2+} , retarding the warping of the active zone sufficiently to prevent the energy barrier for ensuing neighbouring vesicle fusion from increasing fast enough, allowing multiple releases to occur. Our model accounts for the fact that release probability is independent of Rab3A function. □

Methods

Culture experiments. Low-density microisland cultures of hippocampal pyramidal neurons were prepared from E15–E17 wild-type, *rab3A*-mutant and synaptotagmin I mutant mouse embryos as described⁸; five to seven embryos from the same mother were used for every wild-type and *rab3A* mutant culture preparation. Independent cultures of the same genotype yielded reproducible results, and each set of experiments was performed on separate cultures. Some of the experiments were also done on cultures prepared from rat (Fig. 2), and were included in the wild-type data as there were no differences in synaptic properties of the cultures of the two rodent species. Cells were used 8–13 days after plating; coverslips were changed after each recording so that the number of cells/cell pairs given in the figure legends represents the number of coverslips used. extracellular recording solution contained (in mM): NaCl 137, KCl 5, CaCl_2 10, MgCl_2 0.5, D-glucose 10, HEPES–NaOH (pH 7.3) 5, and 100 μM picrotoxin was also added; concentrations of MgCl_2 and CaCl_2 were varied when indicated. The internal recording solution was as before⁸. Recordings were made with axopatch 200 (Axon Instruments). Signals were filtered at 2 kHz, digitized at 5 kHz and analysed with programs written in AXOBASIC, VisualBASIC and C.

Hypertonic solution was applied locally as described⁷, and the quantal release rate was normalized for the number of synapses. The upper limit of the releasable pool size was estimated by the total number of quanta released per synapse during the first 4 s of hypertonic solution application; the lower limit was defined by subtracting the steady-state pool size for the 4-s period. The releasable pool size for wild-type cells determined here was slightly higher than a previous estimate⁷: the discrepancy is probably due to underestimation of the number of synapses in the perfused area, because, depending on the intensity threshold of the fluorescence staining of the synapses (determined visually after immunofluorescence labelling against synapsin I; ref. 8) used for defining synapses, the number of synapses can readily vary by at least 2-fold (C.E.S. and J. F. Wesseling, unpublished observations). The wild-type and mutant culture experiments were nevertheless done under identical conditions; we observed no difference in the pool size among the three groups.

Slice experiments. Acute transverse hippocampal slices were prepared and whole-cell recordings were made as described^{26–28}. Concentrations of CaCl_2 and MgCl_2 were varied as indicated for each experiment. 1 mM QX314 (Sigma) was included in the recording pipette to suppress spike generation in the post-synaptic cell. MK-801 and CNQX (6-cyano-7-nitroquinoxaline-2,3-dione) were from Research Biochemicals International.

Received 28 February; accepted 16 April 1997.

- Südhof, T. C. The synaptic vesicle cycle: a cascade of protein-protein interactions. *Nature* **375**, 645–653 (1995).
- Scheller, R. H. Membrane trafficking in the presynaptic nerve terminal. *Neuron* **14**, 893–897 (1995).
- Pfeffer, S. R. Rab GTPases: master regulators of membrane trafficking. *Curr. Opin. Cell Biol.* **6**, 522–526 (1994).
- Katz, B. *The Release of Neural Transmitter Substances* (Liverpool University Press, Liverpool, 1969).
- Hubbard, J. I. Repetitive stimulation at the mammalian neuromuscular junction, and the mobilization of transmitter. *J. Physiol. (Lond.)* **169**, 641–662 (1963).
- Geppert, M. et al. The role of rab3A in neurotransmitter release. *Nature* **369**, 493–497 (1994).
- Stevens, C. F. & Tsujimoto, T. Estimates for the pool size of releasable quanta at a single central synapse and for the time required to refill the pool. *Proc. Natl Acad. Sci. USA* **92**, 846–849 (1995).

- Geppert, M. et al. Synaptotagmin I: a major Ca^{2+} sensor for transmitter release at a central synapse. *Cell* **79**, 717–727 (1994).
- Rosenmund, C. & Stevens, C. F. Definition of the readily releasable pool of vesicles at hippocampal synapses. *Neuron* **16**, 1197–1207 (1996).
- Mallart, A. & Martin, A. The relation between quantal content and facilitation at the neuromuscular junction of the frog. *R. J. Physiol. (Lond.)* **196**, 593–604 (1968).
- Kamiya, H. & Zucker, R. S. Residual Ca^{2+} and short-term synaptic plasticity. *Nature* **371**, 603–606 (1994).
- Huettnner, J. E. & Bean, B. P. Block of N-methyl-D-aspartate-activated current by the anticonvulsant MK-801: selective binding to open channels. *Proc. Natl Acad. Sci. USA* **85**, 1307–1311 (1988).
- Rosenmund, C., Clements, J. D. & Westbrook, G. L. Non-uniform probability of glutamate release at a hippocampal synapse. *Science* **262**, 754–757 (1993).
- Hessler, N. A., Shirke, A. M. & Manilow, R. The probability of transmitter release at a mammalian central synapse. *Nature* **366**, 569–572 (1993).
- Huang, E. & Stevens, C. F. Estimating the distribution of synaptic reliabilities. *J. Neurophysiol.* (submitted).
- Clements, J. D., Lester, R. A. J., Tong, G., Jahr, C. E. & Westbrook, G. L. The time course of glutamate in the synaptic cleft. *Science* **258**, 1498–1501 (1992).
- Perkel, D. J. & Nicoll, R. A. Evidence for all-or-none regulation of neurotransmitter release: implications for long-term potentiation. *J. Physiol. (Lond.)* **471**, 481–500 (1993).
- Olverman, H. J., Jones, A. W., Mewett, K. N. & Watkins, J. C. Structure/activity relations of N-methyl-D-aspartate receptor ligands as studied by their inhibition of [^3H]D-2-amino-5-phosphonopentanoic acid binding in rat brain membranes. *Neuroscience* **26**, 17–31 (1988).
- Holz, R. W., Brondyk, W. H., Senter, R. A., Kuizon, L. & Macara, I. G. Evidence for the involvement of Rab3A in Ca^{2+} -dependent exocytosis from adrenal chromaffin cells. *J. Biol. Chem.* **269**, 10229–10234 (1994).
- Johannes, L. et al. The GTPase rab3A negatively controls calcium-dependent exocytosis in neuroendocrine cells. *EMBO J.* **13**, 2029–2037 (1994).
- Stahl, B., Chou, J. H., Li, C., Südhof, T. C. & Jahn, R. Rab3 reversibly recruits rabphilin to synaptic vesicles by a mechanism analogous to raf recruitment by ras. *EMBO J.* **15**, 1799–1809 (1996).
- Mallart, A. & Martin, A. R. An analysis of facilitation of transmitter release at the neuromuscular junction of the frog. *J. Physiol. (Lond.)* **193**, 593–604 (1967).
- Manabe, T., Wyllie, D. J. A., Perkel, D. J. & Nicoll, R. A. Modulation of synaptic transmission and long-term potentiation: effects of paired pulse facilitation and EPSC variance in the CA1 region of the hippocampus. *J. Neurophysiol.* **70**, 1451–1459 (1993).
- Schulz, P. E., Cook, E. P. & Johnston, D. Changes in paired-pulse facilitation suggest presynaptic involvement in long-term potentiation. *J. Neurosci.* **14**, 5325–5337 (1994).
- Raastad, M., Storm, J. F. & Andersen, P. Putative single quantum and single fibre excitatory postsynaptic currents show similar amplitude range and variability in rat hippocampal slices. *Eur. J. Neurosci.* **4**, 113–117 (1992).
- Allen, C. & Stevens, C. F. An evaluation of causes for unreliability of synaptic transmission. *Proc. Natl Acad. Sci. USA* **91**, 10380–10383 (1994).
- Stevens, C. F. & Wang, Y. Changes in reliability of synaptic function as a mechanism for plasticity. *Nature* **371**, 704–707 (1994).
- Stevens, C. F. & Wang, Y. Facilitation and depression at single central synapses. *Neuron* **14**, 795–802 (1995).
- Triller, A. & Korn, H. Transmission at a central inhibitory synapse. III. Ultrastructure of physiologically identified and stained terminals. *J. Neurophysiol.* **48**, 708–736 (1982).
- Rybin, V. et al. GTPase activity of Rab5 acts as a timer for endocytic membrane fusion. *Nature* **383**, 266–269 (1996).

Acknowledgements. The authors are listed in alphabetical order. We thank J. Wesseling for discussions, and C. Boyer for help in the culture preparation. This work was supported by the Howard Hughes Medical Institute (C.E.S. and T.C.S.), the NIH (C.E.S.) and the National Alliance for Research on Schizophrenia and Depression (Y.G.).

Correspondence should be addressed to Y.G. (e-mail: ygoda@jeeves.ucsd.edu).

C-terminal binding domain of Rho GDP-dissociation inhibitor directs N-terminal inhibitory peptide to GTPases

Yuying Q. Gosser*, Tyzoon K. Nomanbhoy†, Behzad Aghazadeh*‡, Danny Manor†, Carolyn Combs†, Richard A. Cerione† & Michael K. Rosen*

* Cellular Biochemistry and Biophysics Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

† Departments of Biochemistry, Molecular and Cell Biology, and Department of Pharmacology, Cornell University, Ithaca, New York 14853, USA

‡ Graduate Program in Physiology and Biophysics, Cornell University Graduate School of Medical Sciences, New York, New York 10021, USA

The Rho GDP-dissociation inhibitors (GDIs) negatively regulate Rho-family GTPases^{1,2}. The inhibitory activity of GDI derives both from an ability to bind the carboxy-terminal isoprene of Rho family members and extract them from membranes^{3,4}, and from inhibition of GTPase cycling between the GTP- and GDP-bound states^{4,5}. Here we demonstrate that these binding and inhibitory functions of rhoGDI can be attributed to two structurally

distinct regions of the protein. A carboxy-terminal folded domain of relative molecular mass 16,000 (M_r 16K) binds strongly to the Rho-family member Cdc42, yet has little effect on the rate of nucleotide dissociation from the GTPase. The solution structure of this domain shows a β -sandwich motif with a narrow hydrophobic cleft that binds isoprenes, and an exposed surface that interacts with the protein portion of Cdc42. The amino-terminal region of rhoGDI is unstructured in the absence of target and contributes little to binding, but is necessary to inhibit nucleotide dissociation from Cdc42. These results lead to a model of rhoGDI function in which the carboxy-terminal binding domain targets the amino-terminal inhibitory region to GTPases, resulting in membrane extraction and inhibition of nucleotide cycling.

In an attempt to understand the molecular mechanisms of the inhibitory activities of GDI, we have performed a series of structural and biochemical analyses of the archetypal protein in this family, rhoGDI. Investigations focused on two carboxy-terminal fragments of bovine rhoGDI, GDI Δ 22 (residues 23–204) and GDI Δ 59 (residues 60–204). GDI Δ 22 is similar to a construct (residues 25–204) previously reported to be fully functional in both membrane-extraction and nucleotide-dissociation assays⁶, and is therefore used here as a model of the full-length protein. ¹⁵N-labelled samples of GDI Δ 22 and GDI Δ 59 were analysed by ¹H/¹⁵N heteronuclear single quantum coherence (HSQC) nuclear magnetic resonance (NMR) spectra⁷ (Fig. 1a,b). The wide ¹HN chemical shift dispersion (6.5–10.4 p.p.m.) and relatively uniform peak intensity in Fig. 1b suggest that GDI Δ 59 forms a single folded

domain in solution. In contrast, whereas most resonances in the GDI Δ 22 spectrum show relatively weak intensity and are coincident with peaks observed for GDI Δ 59, a cluster of 40 very intense peaks is also observed near the centre of Fig. 1a that is not seen in Fig. 1b. Backbone sequential assignment of both constructs confirmed that these intense peaks represent residues 23–64 of GDI Δ 22 (residue numbers throughout are referenced to full-length rhoGDI). The poor amide chemical-shift dispersion (¹H range, 8–8.6 p.p.m.) and decrease in intensity of these peaks after presaturation of solvent (data not shown), coupled with long amide nitrogen transverse relaxation times (200 ± 50 ms, compared with 50 ± 10 ms for the well-dispersed amides) indicate that the N-terminal residues of GDI Δ 22 are disordered in solution. The correspondence between the peaks for GDI Δ 22 (Fig. 1a) and GDI Δ 59 (Fig. 1b) (average absolute value of ¹H and ¹⁵N chemical shift deviations for residues 69–204: 0.006 ± 0.004 and 0.03 ± 0.04 p.p.m., respectively) indicates that the C-terminal domain is structurally independent from this N-terminal peptide, and is conformationally identical in the two constructs.

In order to understand the functions of the N- and C-terminal regions of rhoGDI, the affinities of the full-length protein, GDI Δ 22, GDI Δ 42 (residues 43–204) and GDI Δ 59 for isoprenylated Cdc42 were measured by fluorescence resonance energy transfer between the GTPase (complexed with the fluorescent nucleotide 2'(3')-O-(N-methylanthraniloyl)GDP (Mant-GDP)) and the GDI constructs (covalently modified with fluorescein maleimide). The affinity of the constructs for Cdc42 decreases as increasing amounts of the N

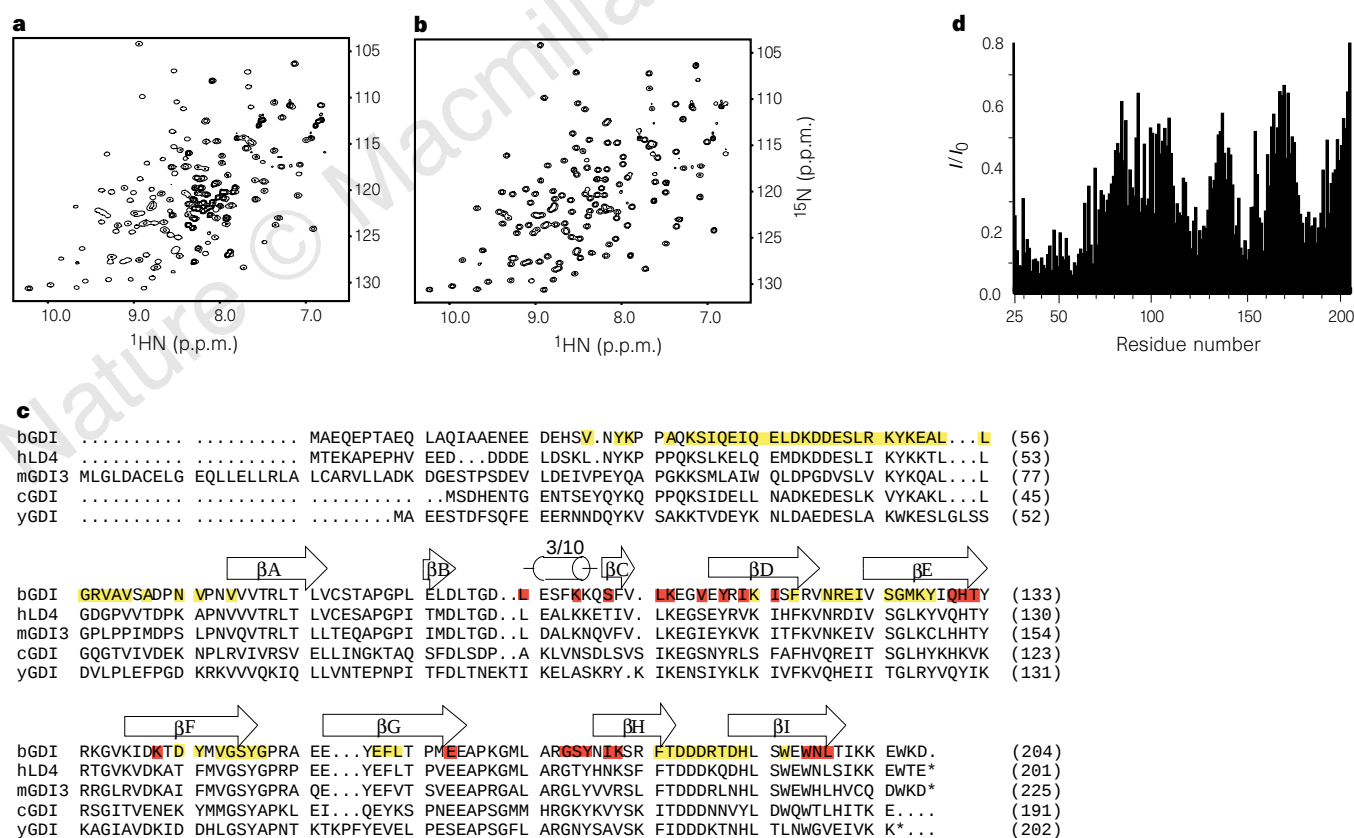


Figure 1 **a**, ¹H/¹⁵N HSQC spectra of GDI Δ 22 (**a**) and GDI Δ 59(**b**). **c**, Sequence alignment of rhoGDI proteins: bovine RhoGDI (bGDI), human D4 (hLD4), murine GDI3 (mGDI3), *Caenorhabditis elegans* GDI (cGDI) and *Saccharomyces cerevisiae* GDI (yGDI). The secondary structure²⁷ of GDI Δ 59 is indicated above the alignment. Residues affected by addition of acetyl-I-KKSRRC(S-farnesyl) peptide (see text) are red. Residues with ¹H/¹⁵N HSQC peak intensities decreased to <25% of their original values, or showing ¹HN chemical shift

changes >0.05 p.p.m. on addition of 0.9 equivalents of non-isoprenylated Cdc42(GDP) are yellow. **d**, Normalized ¹H/¹⁵N HSQC peak intensity, $I_{0.9}/I_{0.0}$, versus residue number in GDI Δ 22 (wild-type rhoGDI numbering). $I_{0.0}$ and $I_{0.9}$ are intensity measured for the free protein, and after addition of 0.9 equivalents non-isoprenylated Cdc42(GDP), respectively. Prolines and residues 26, 29, 30, 80, 107, 135, 153 and 169, which had amide peaks that were either badly overlapped or too weak for intensity to be measured, are not included.

terminus are removed (Fig. 2a). However, even GDIA59, which contains almost exclusively the folded domain of rhoGDI, still binds the GTPase in the high nanomolar range. In contrast, NMR studies with an N-terminal rhoGDI peptide (residues 23–58) show extremely low affinity ($K_D > 5$ mM) for Cdc42 (data not shown). Both domains of the regulator are required to partition Cdc42 from membranes, however, as GDIA42 and GDIA59 are inactive in membrane extraction assays (data not shown).

The different rhoGDI constructs were also examined for their ability to inhibit nucleotide dissociation from isoprenylated Cdc42. Although the GTPase is nearly saturated with regulator in all cases (93–98% based on the K_D in Fig. 2a; see Methods), GDP dissociation is slowed only in the presence of full-length rhoGDI and GDIA22 (Fig. 2b). Identical results were seen in assays performed at 1 μ M Cdc42 and 400 μ M GDIA59, strongly suggesting that the differences in activity are not simply a result of differences in binding affinity. Thus inhibitory activity resides in residues 23–42, with C-terminal residues primarily contributing binding energy rather than inhibitory effect. The combined NMR and biochemical data provide strong support for a model in which rhoGDI consists of a folded C-terminal binding domain that targets an inhibitory N-terminal peptide to GTPases.

In an attempt to develop a more detailed model of the rhoGDI–GTPase complex, we determined by NMR spectroscopy the three-dimensional solution structure of GDIA59. The structure was calculated using a hybrid distance geometry/simulated annealing protocol using a total of 1,992 distance restraints, and 110 ϕ , 52 ψ and 30 χ_1 torsional restraints. The structures agree well with the experimental data, and show good covalent geometry and non-bonded contacts (see Methods). A stereo view of a backbone superposition of residues 69–202 of the 20 lowest-energy structures of GDIA59 is shown in Fig. 3a. Residues 60–68 and 203–204 show no long-range nuclear Overhauser enhancements (NOEs), are disordered in the structure, and have been removed for clarity. The structure is well defined by the NMR data, with atomic root-mean-squared deviation (r.m.s.d.) from the mean coordinates of 0.64 ± 0.06 Å for backbone atoms and 1.18 ± 0.09 Å for all heavy atoms in residues 69–202.

The structure of GDIA59 consists of two largely antiparallel β -sheets that pack against each other in a parallel manner, forming a β -sandwich. The protein has an immunoglobulin-like fold, with a modified Greek-key β -strand topology. The top sheet in Fig. 3b consists of the antiparallel strands F, E, H and I, flanked by strand C, which aligns parallel to strand I (see Fig. 1c). The bottom sheet contains four antiparallel strands, B, A, D and G. The β -sandwich is closed at the top and bottom by loops, and is capped at one end by a short 3_{10} helix, and side chains from residues in strands B and C. The opposing end is open to solvent, and shows a deep pocket between the two sheets. The disordered N terminus extends from the top surface of the protein in Fig. 3b near loops H–I and D–E. The structure shows little resemblance to RabGDI, which selectively binds prenylated GDP-bound Rab GTPases and is involved in vesicle transport⁸. This discrepancy is not unexpected, given the functional differences between the two GDI families, including sensitivity to GTPase nucleotide state and isoprenylation patterns⁸.

The minimized average NMR structure of GDIA59 is shown in Fig. 3c, viewed into the pocket at the open end of the β -sandwich. Its base and inner walls are lined by several conserved hydrophobic residues, including Leu 104, Tyr 110, Ile 112, Thr 132, Ala 165, Pro 166, Tyr 175, Ile 177, Trp 194, Leu 196 and Ile 198. The side chain of a single conserved hydrophilic amino acid, Gln 130, extends into the pocket, and may form a hydrogen bond with the nearby hydroxyl group of Tyr 128. Two lines of evidence indicate that this is the isoprene binding site of rhoGDI. First, addition of the farnesylated peptide acetyl-KKSRRC(S-farnesyl) to GDIA59 as a readily available mimic of the isoprenylated C terminus of Rho-family GTPases (note that RhoB, which is farnesylated, is a target of rhoGDI (ref. 3)

and mGDI3 (ref. 9)) caused large, saturable changes to the backbone and/or side-chain amide (^1H , >0.1 , or ^{15}N , >1.0 p.p.m.) or ^{13}C methyl (>0.4 p.p.m.) resonances of many residues in and near the pocket (Fig. 1c). No changes were observed when non-isoprenylated peptide was added. However, a nearly identical set of residues was affected, although to a lesser degree, by addition of *N*-acetyl-S-farnesyl cysteine to the limit of its aqueous solubility. Second, it has been shown that the 20-fold difference in affinity for Cdc42 between rhoGDI and its homologue D4 (Fig. 1c) is due to a difference of isoleucine (rhoGDI) versus asparagine (D4) at residue 177 (ref. 6), which lines one side of the pocket.

In most of the structures in the final ensemble, the isoprene-binding site is approximately 4.5–5.5 Å wide, 9 Å long and 10 Å deep. The pocket is reminiscent of that found in fatty-acid binding proteins (FABPs), which sequester linear hydrophobic ligands in a cavity between two sheets¹⁰. However, the FABP pocket is much larger and contains many more polar and charged residues than that of rhoGDI, allowing promiscuous binding to a variety of ligands in many different conformations. A distinct feature of prenyl chains is their conformational restriction by local steric interactions, including allylic 1,3 ($\text{A}^{1,3}$) and pseudo-1,3 diaxial (*gauche*⁺, *gauche*[−]) strain^{11,12}. The hydrophobicity and small size of the binding pocket, combined with the geometric restrictions of the isoprene, probably impart selectivity for prenyl groups to rhoGDI.

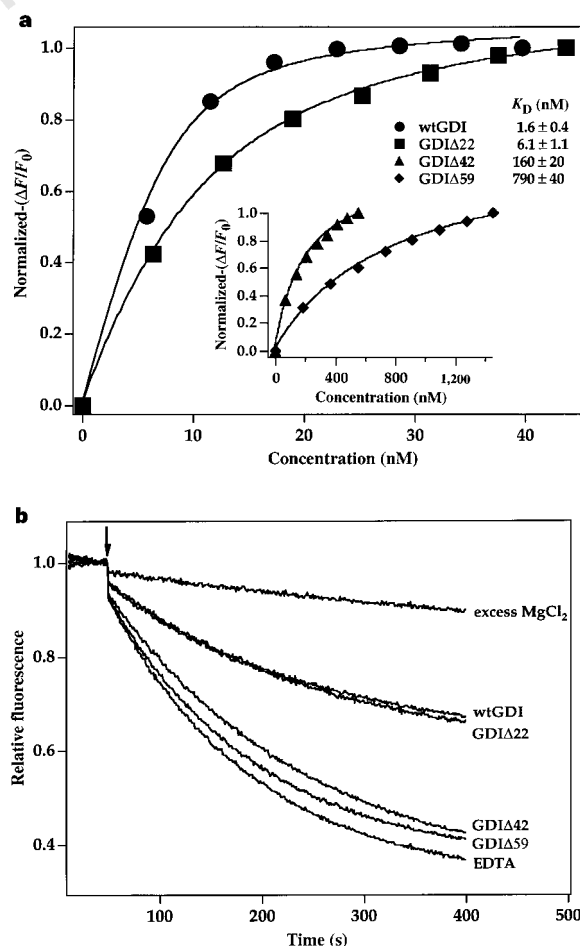
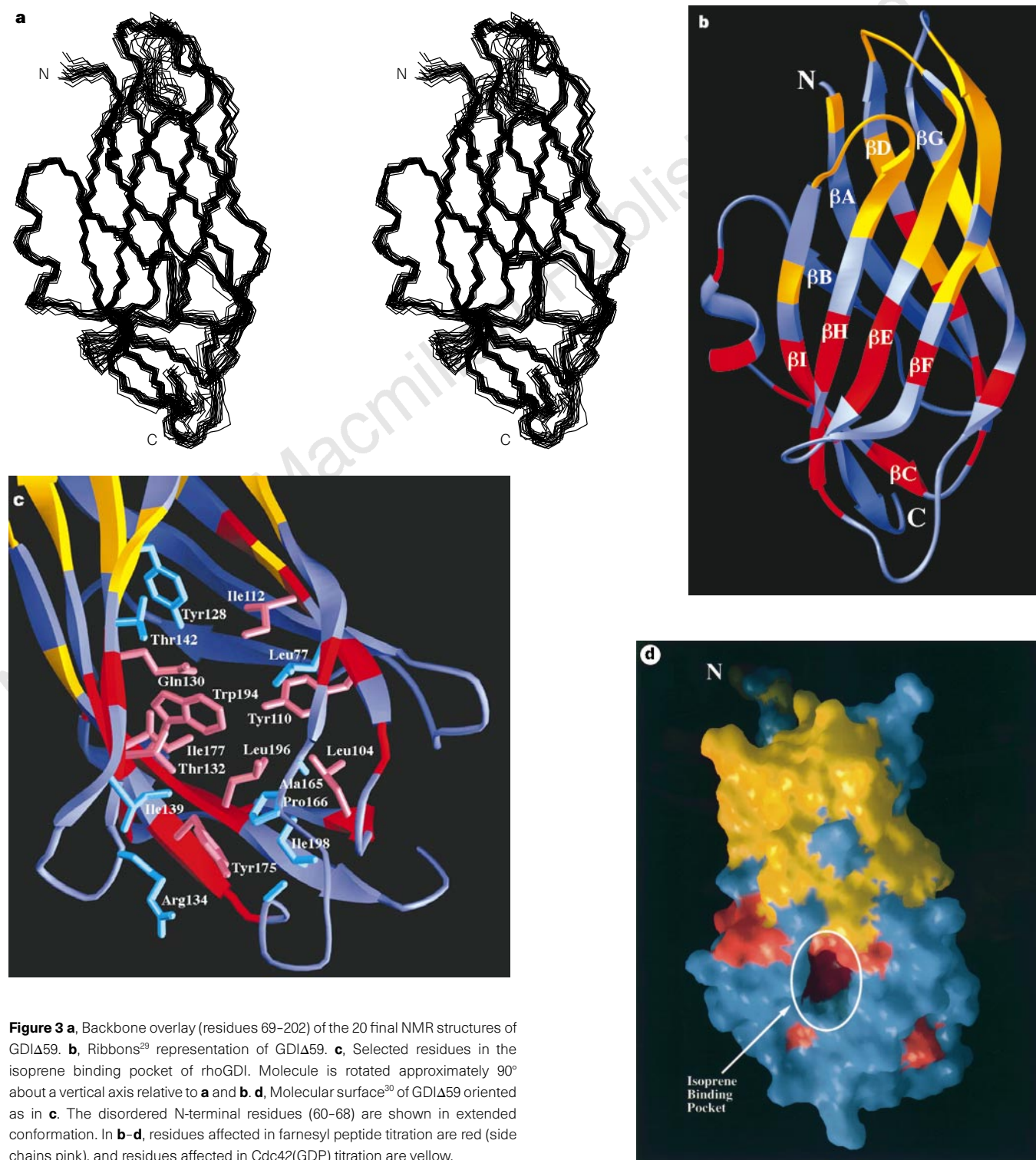


Figure 2 **a**, Normalized change in fluorescence (F) of the (Mant-GDP)–Cdc42 complex on addition of fluorescein-conjugated GDI constructs. Binding affinity (K_D) was calculated as described¹⁶; wtGDI, wild-type GDI. **b**, Dissociation of Mant-GDP (indicated by decreased fluorescence²⁸) from isoprenylated Cdc42 following addition of EDTA (arrow) in the presence of nearly saturating amounts of GDI proteins.

The specificity of GDIs for Rho-family members indicates that, although the isoprene is important for binding, other protein-protein contacts are necessary for high-affinity interactions. To identify these regions, ^{15}N -GDI Δ 22 was titrated with unlabelled, non-isoprenylated Cdc42 loaded with GDP. At subsaturating GTPase levels, differential exchange broadening (as manifested by decreased peak intensity) in $^1\text{H}/^{15}\text{N}$ -HSQC spectra can be used to identify residues with chemical shifts that are most perturbed by binding. A plot of peak intensity at 0.9 equivalents of Cdc42 (40–50% saturation based on peak volumes for four resonances that show distinct bound and free species) versus amino-acid sequence

of GDI Δ 22 is shown in Fig. 1d. N-terminal residues 23–63 are strongly affected. Four additional regions of the folded C-terminal domain, spanning the D–E and H–I loops at the apex of the molecule, and adjacent parts of strands H, E, F and G, are also affected. Most of the residues in these regions, in particular 119–128, 147–150 and 181–186, are conserved throughout the GDI family. The protein-binding residues form a continuous surface immediately adjacent to the isoprene-binding site of rhoGDI (Fig. 3d). The N-terminal inhibitory peptide extends from strand A between the D–E and H–I loops at the back of this surface, where it could easily contact the bound GTPase to impart GDI activity.



The combined structural and biochemical data suggest a model for the full rhoGDI–Cdc42 complex in which the GTPase sits on top of the C-terminal domain of the regulator, with its attached isoprene buried in the GDI hydrophobic pocket. These interactions anchor the inhibitory N-terminal peptide to the GTPase, allowing it to block exchange of GDP, and also probably binding of GAPs and GNRFs. The combined interactions of the GDI N and C termini then provide sufficient thermodynamic driving force to compete GTPases away from membranes.

We observed that GDIΔ42, which binds Cdc42 only 100-fold more weakly than full-length protein, is unable to extract the GTPase from membranes. Also, the binding energy of rhoGDI for Cdc42 ($\sim 12 \text{ kcal mol}^{-1}$) is very similar to the membrane–water partitioning energy reported¹³ for model geranylgeranylated peptides ($\sim 12 \text{ kcal mol}^{-1}$). These data suggest that there is a delicate thermodynamic balance in the partitioning of rhoGTPases between membranes and rhoGDI family members in the cell. In this case, small changes in GTPase affinity for either species could result in large changes in distribution between membranes and the cytoplasm. Carboxymethylation of Rac2 in neutrophils in response to various signals correlates with migration of the GTPase from a rhoGDI complex in the cytoplasm to the plasma membrane¹⁴. This modification, which is known to increase the partitioning of prenyl peptides into negatively charged liposomes by a factor of at least 10–20 (refs 13, 15), could thus be used in the cell to release GTPases from inhibitory GDIs by tipping the thermodynamic balance in favour of membranes.

It has been reported that in T cells the N-terminal region of the rhoGDI homologue D4 is specifically cleaved at two positions (residues 18–19 and 54–55) by two different ICE-like proteases following induction of apoptotic signalling pathways^{16,17}. The larger of the two resulting C-terminal proteolytic fragments is likely to retain inhibitory activity, but the smaller will be unable to bind GTPases in the environment of the cell. Our finding that the N terminus of rhoGDI is disordered in solution suggests that this region of all GDI family members will be proteolytically sensitive. Thus proteolysis may be a general mechanism for regulating the activity of these proteins. Moreover, given that GDIΔ22 retains membrane extraction and nucleotide-exchange inhibition activity, but that GDIΔ42 and GDIΔ59 do not, and that the N-terminal extension of mGDI3 (Fig. 1c) seems to mediate membrane localization of this protein (R.A.C., unpublished observations), it is possible that proteolysis at different sites in the GDI N terminus could lead to fragments with different inhibitory properties. Our data now provide a structural and thermodynamic framework for continued analysis of the regulatory functions of GDI proteins in the cell.

Note added in Proof: Preliminary NOE analyses of a GDIΔ59–farnesyl peptide complex suggest that the isoprene binding pocket may deepen on ligation through structural changes that allow it to merge with a small adjacent cavity seen in some members of the free protein ensemble. □

Methods

Protein expression and purification. ^{15}N labelled or $^{15}\text{N}/^{13}\text{C}$ labelled GDIΔ22 and GDIΔ59 were overexpressed in *Escherichia coli* strain BL21(DE3) grown in minimal media containing $^{15}\text{NH}_4\text{Cl}$ or $^{15}\text{NH}_4\text{Cl}/^{13}\text{C}_6\text{-glucose}$, respectively, and purified chromatographically. NMR samples were 2 mM GDI protein in 20 mM sodium phosphate buffer, pH 6.8, except for titrations where protein concentration was 0.5 mM. Non-isoprenylated and isoprenylated Cdc42 were expressed in bacteria and insect cells, respectively, and purified^{18,19}.

GDI binding assay. GDI proteins were conjugated at Cys 79 with fluorescein by reaction with a 10-fold excess of fluorescein maleimide for 1 h at room temperature in the dark. Cdc42 was loaded with Mant-GDP as described¹⁸. Energy transfer between the two chromophores in the GDI–GTPase complex results in quenching of Mant fluorescence ($\lambda_{\text{excitation}}$ 350 nm; $\lambda_{\text{emission}}$ 440 nm).

Nucleotide dissociation assay. Cdc42 (25 nM) prelabelled with Mant-GDP

was incubated in the presence of 125 mM GDP and either 100 nM wild-type rhoGDI, 250 nM GDIΔ22, 2 μM GDIΔ42, 10 μM GDIΔ59, in the absence of any GDI construct, or in 10 mM MgCl_2 . Nucleotide dissociation was initiated by addition of EDTA to 6 mM. Mant fluorescence was monitored as above.

NMR spectroscopy. NMR spectra were collected at 30 °C on a Varian UNITY+ 600 MHz spectrometer. Backbone sequential assignments of GDIΔ59 and GDIΔ22 were completed using standard methods through HNCO, HNCA, Cb, CbCoNHH, HNCAHa, HCC-TOCSY-NNH and ^{15}N -edited TOCSY spectra²¹. Side-chain assignments were completed through analysis of CCC-TOCSY-NNH, HCCH-TOCSY and HCCH-COSY spectra²¹. Stereospecific assignments of 9 Val and 11 Leu methyl groups were obtained from constant time- $^1\text{H}/^{13}\text{C}$ HSQC data acquired on a 10% ^{13}C -labelled sample²².

Structure determination. Structures were determined through a standard hybrid distance geometry/simulated annealing protocol implemented in X-PLOR²³. Distance restraints were obtained from three-dimensional ^{15}N , simultaneous $^{15}\text{N}/^{13}\text{C}$ - and ^{13}C -edited NOESY spectra⁷, and classified into four categories: 1.8–2.7 Å (2.9 Å for amides), 1.8–3.3 Å (3.5 Å for amides), 1.8–5.0 Å and 1.8–6.0 Å, based on peak intensities in spectra recorded with mixing times of 50 or 60 ms. A further 98 restraints were included for 49 slowly exchanging backbone amides (1.5–2.8 Å (H–O) and 2.4–3.5 Å (N–O)). A total of 110 ϕ (± 30 – 50°) and 52 ψ ($\pm 50^\circ$ for helix, $\pm 90^\circ$ for sheets) torsional restraints were based on $^3J_{\text{HN-H}\alpha}$ coupling constants measured in an HNHA experiment²⁴, and on the correlation between $\text{C}\alpha$ and $\text{C}\beta$ chemical shifts and backbone torsion angles²⁵, respectively. At the final stages of the calculations, 30 χ^1 restraints ($\pm 30^\circ$) were added for Val, Ile, Thr, Leu and aromatic residues based on intra-residue and sequential NOEs²⁶. For the final ensemble of 20 structures, the r.m.s.d. values from experimental restraints are: NOE, 0.017 ± 0.001 Å; torsional, $0.22 \pm 0.07^\circ$. No structures showed distance violations greater than 0.3 Å or dihedral violations greater than 5° . X-PLOR energies are (kcal mol^{-1}) E_{NOE} , 30 ± 5 (force constant, $50 \text{ kcal mol}^{-1} \text{Å}^{-2}$); E_{cdih} , 0.6 ± 0.4 (force constant, $200 \text{ kcal mol}^{-1} \text{rad}^{-2}$); E_{total} , 349 ± 14 ; E_{repeb} , 63 ± 5 . A quartic repulsion potential (force constant, $4 \text{ kcal mol}^{-1} \text{Å}^{-4}$) was used with atomic radii set to 0.80 times their actual values. R.m.s.d. values from ideal geometries are: bonds, 0.00 ± 0.00 Å; angles, $0.56 \pm 0.01^\circ$; impropers, $0.38 \pm 0.01^\circ$.

Received 8 April; accepted 16 May 1997.

- Ueda, T., Kikuchi, A., Ohga, N., Yamamoto, J. & Takai, Y. Purification and characterization from bovine brain cytosol of a novel regulatory protein inhibiting the dissociation of GDP from and the subsequent binding of GTP to *rhoB* p20, a *ras* p21-like GTP-binding protein. *J. Biol. Chem.* **265**, 9373–9380 (1990).
- Fukumoto, Y. *et al.* Molecular cloning and characterization of a novel type of regulatory protein (GDI) for the *rho* proteins, *ras* p21-like small GTP-binding proteins. *Oncogene* **5**, 1321–1328 (1990).
- Isomura, M., Kikuchi, A., Ohga, N. & Takai, Y. Regulation of binding of *rhoB* p20 to membranes by its specific regulatory protein, GDP dissociation inhibitor. *Oncogene* **6**, 119–124 (1991).
- Bokoch, G. M. & Der, C. J. Emerging concepts in the Ras superfamily of GTP-binding proteins. *FASEB J.* **7**, 750–759 (1993).
- Hancock, J. F. & Hall, A. A novel role for RhoGDI as an inhibitor of GAP proteins. *EMBO J.* **12**, 1915–1921 (1993).
- Platko, J. V. *et al.* A single residue can modify target-binding affinity and activity of the functional domain of the Rho-subfamily GDP dissociation inhibitors. *Proc. Natl Acad. Sci. USA* **92**, 2974–2978 (1995).
- Kay, L. E. Field gradient techniques in NMR spectroscopy. *Curr. Opin. Struct. Biol.* **5**, 674–681 (1995).
- Schalk, I. *et al.* Structure and mutational analysis of Rab GDP-dissociation inhibitor. *Nature* **381**, 42–48 (1996).
- Zalcman, G. *et al.* RhoGDI-3 is a new GDP dissociation inhibitor (GDI). *J. Biol. Chem.* **271**, 30366–30374 (1996).
- Storch, J. Diversity of fatty acid-binding protein structure and function: studies with fluorescent ligands. *Mol. Cell. Biochem.* **123**, 45–53 (1993).
- Hoffmann, R. W. Allylic 1,3-strain as a controlling factor in stereoselective transformations. *Chem. Rev.* **89**, 1841–1860 (1989).
- Wiberg, K. B. & Murko, M. A. Rotational barriers. 2. Energies of alkane rotamers. An examination of gauche interactions. *J. Am. Chem. Soc.* **110**, 8029–8038 (1988).
- Silvius, J. R. & l'Heureux, F. Fluorimetric evaluation of the affinities of isoprenylated peptides for lipid bilayers. *Biochemistry* **33**, 3014–3022 (1994).
- Philips, M. R. *et al.* Carboxyl methylation of Ras-related proteins during signal transduction in neutrophils. *Science* **259**, 977–980 (1993).
- Ghomashchi, R., Zhang, X., Liu, L. & Gelb, M. H. Binding of prenylated and polybasic peptides to membranes: affinities and interside exchange. *Biochemistry* **34**, 11910–11918 (1995).
- Na, S. *et al.* D4-GDI, a substrate of CPP32, is proteolyzed during Fas-induced apoptosis. *J. Biol. Chem.* **271**, 11209–11213 (1996).
- Danley, D. E., Chuang, T.-H. & Bokach, G. M. Defective Rho GTPase regulation by IL-1 β -converting enzyme-mediated cleavage of D4 GDP dissociation inhibitor. *J. Immunol.* **157**, 500–503 (1996).
- Nomanbhoy, T. K. & Cerione, R. A. Characterization of the interaction between RhoGDI and Cdc42Hs using fluorescence spectroscopy. *J. Biol. Chem.* **271**, 10004–10009 (1996).
- Wu, J. W., Leonard, D. A., Cerione, R. A. & Manor, D. Interaction between Cdc42Hs and Rho-GDI is mediated through the Rho-insert region. *J. Biol. Chem.* (in the press).
- Clore, G. M. & Gronenborn, A. M. Multidimensional heteronuclear nuclear magnetic resonance of proteins. *Methods Enzymol.* **239**, 349–363 (1994).

21. Edison, A. S., Abildgaard, E., Westler, W. M., Mooberry, E. S. & Markley, J. L. Practical introduction to theory and implementation of multinuclear, multidimensional nuclear magnetic resonance experiments. *Methods Enzymol.* **239**, 3–79 (1994).
22. Neri, D., Szyperki, T., Otting, G., Senn, H. & Wuthrich, K. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry* **28**, 7510–7516 (1989).
23. Brünger, A. T. *X-PLOR manual* (Yale Univ. Press, New Haven, CT, 1993).
24. Kuboniwa, H., Grzesiek, S., Delaglio, F. & Bax, A. Measurement of HN-H α couplings in calcium-free calmodulin using new 2D and 3D water-flip-back methods. *J. Biomol. NMR* **4**, 871–878 (1994).
25. Spera, S. & Bax, A. Empirical correlation between protein backbone conformation and C α and C β ^{13}C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* **117**, 5491–5495 (1991).
26. Nilges, M., Clore, G. M. & Gronenborn, A. M. ^1H -NMR stereospecific assignments by conformational data-base searches. *Biopolymers* **29**, 813–822 (1990).
27. Laskowski, R. A., Rullman, J. A. C., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–496 (1996).
28. Leonard, D. A., Evans, T., Hart, M., Cerione, R. A. & Manor, D. Investigation of the GTP-binding/GTPase cycle of Cdc42Hs using fluorescence spectroscopy. *Biochemistry* **33**, 12323–12328 (1994).
29. Carson, M. J. Ribbons 2.0. *J. Appl. Crystallogr.* **24**, 958–961 (1991).
30. Nichols, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Protein Struct. Funct. Genet.* **11**, 281–296 (1990).

Acknowledgements. We thank J. Kelly for participating in the sequential assignment of GDIA59; L. E. Kay for discussion and for providing the NMR pulse sequences; B. A. Johnson and F. Delaglio for data analysis and processing software; D. Live for assistance with NMR data acquisition; J. Hubbard for computer system support and assistance with figure preparation; and Y. M. Chook and L. E. Kay for critical reading of the manuscript. This work was supported by a grant from the Society of the Memorial Sloan-Kettering Cancer Center.

Correspondence and requests for materials should be addressed to M.K.R. (e-mail: rosen@mrnmr1.ski.mskcc.org). Coordinates of the minimized average GDIA59 structure and of the final ensemble of 20 structures have been deposited in the Brookhaven Protein Data Bank (accession numbers 1gdf and 1ajw, respectively).

Synergistic activation of transcription by CBP and p53

Wei Gu, Xiao-Lu Shi* & Robert G. Roeder

Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

The tumour suppressor p53 is a transcriptional regulator whose ability to inhibit cell growth is dependent upon its transactivation function^{1–3}. Here we demonstrate that the transcription factor CBP, which is also implicated in cell proliferation and differentiation^{4–14}, acts as a p53 coactivator and potentiates its transcriptional activity. The amino-terminal activation domain of p53 interacts with the carboxy-terminal portion of the CBP protein both *in vitro* and *in vivo*. In transfected SaoS-2 cells, CBP potentiates activation of the *mdm-2* gene by p53 and, reciprocally, p53 potentiates activation of a Gal4-responsive target gene by a Gal4(1–147)–CBP(1678–2441) fusion protein. A double point mutation that destroys the transactivation function of p53 also abolishes its binding to CBP and its synergistic function with CBP. The ability of p53 to interact physically and functionally with a co-activator (CBP) that has histone acetyltransferase activity^{15,16} and with components (TAFs)^{17,18} of the general transcription machinery indicates that it may have different functions in a multistep activation pathway.

CBP functions as a transcriptional coactivator through interactions with a number of cellular activators^{4–14} and, possibly, with components of the basal transcription machinery⁶. CBP has at least three domains that bind transcription factors: a region containing residues 451–662 (designated CBP1) which is required for interactions with CREB (refs 6, 7), c-Jun (ref. 7) and c-Myb (ref. 11); a region containing residues 1,680–1,891 (designated CBP2) which is essential for interactions with E1a (refs 8, 9), c-Fos (ref. 10) and TFIIB (ref. 6); and a carboxy-terminal region containing residues 1,990–2,441 (designated CBP3) which is important for interactions with SRC-1 (ref. 12) and Jun-B (ref. 14) and which contains a

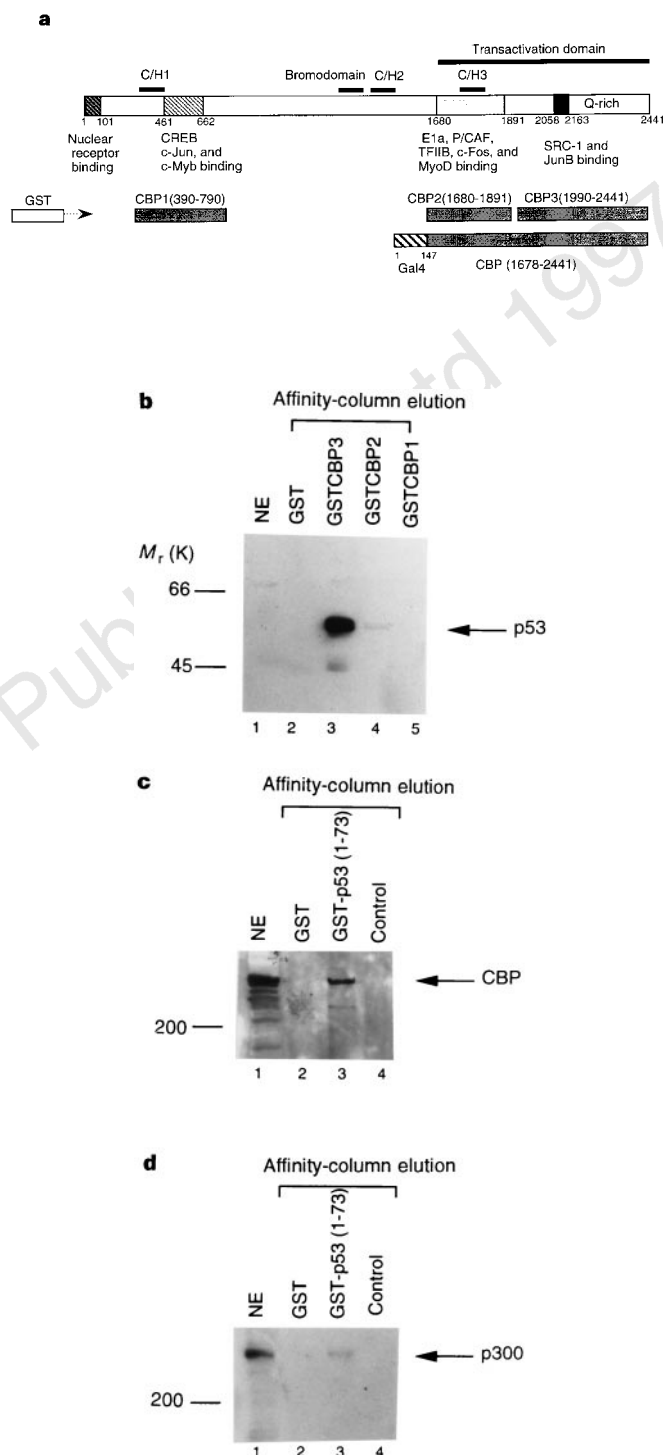


Figure 1 p53 association with CBP. **a**, Schematic representations of CBP and its functional/protein interaction domains, the hybrid activator Gal4–CBP(1678–2441), and the GST–CBP fusion proteins used for affinity chromatography. C/H: cysteine/histidine-rich region. **b**, Affinity purification of p53 on immobilized GST–CBP3. Western blot analysis with anti-p53 (antibody DO-1) of input HeLa nuclear extract (0.5% of 4 mg total protein) and the total 1 M KCl eluates from the indicated GST–CBP affinity columns. **c**, **d**, Transactivation domain of p53 interacts with CBP and p300, respectively. Western blot analysis with anti-CBP3 (**c**) or anti-p300 (**d**) of input HeLa nuclear extract (5% of 1 mg total protein), and half of the total eluates from the GST, GST–p53 affinity columns and from the control column (the GST–p53 column without HeLa nuclear extract loading).

* Present address: HHMI/Pharmacology Department, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, Texas 75235-9050, USA.

transactivation domain⁶. These three domains were fused to GST (Fig. 1a), expressed in *Escherichia coli*, and used as ligands for affinity chromatography¹⁹. To search for proteins that interact with CBP, the corresponding GST–CBP affinity columns and a GST control column were loaded with equal amounts of HeLa nuclear extract. Bound proteins were then eluted with high salt, resolved by SDS–PAGE, and analysed by silver staining and western blot analysis. This analysis (Fig. 1b) identified the tumour suppressor p53 as one of the CBP-interacting proteins in nuclear extracts. Thus, whereas p53 was barely detectable in a crude nuclear extract (Fig. 1b, lane 1), consistent with its low level of expression in HeLa cells²⁰, it was enriched in the high-salt fraction from the GST–CBP3 column (Fig. 1b, lane 3). As relatively little or no p53 was detected in the eluates from the GST, GST–CBP1 or GST–CBP2 columns (Fig. 1b), this result suggests that p53 specifically binds to GST–CBP3 with high affinity.

This discovery of a CBP–p53 interaction was corroborated by a reciprocal experiment in which we investigated the binding of nuclear extract proteins to a GST–p53 affinity column on which the transactivation domain (residues 1–73) of p53 protein fused to GST served as a ligand. The 265K CBP was detected in the high-salt eluate from the GST–p53 column, but not in that from the GST column (Fig. 1c). The CBP-related protein p300 (refs 8, 9) also bound to GST–p53 in the same experiment (Fig. 1d), indicating that there are specific interactions between the transactivation domain of p53 and CBP/p300.

To determine whether p53 directly interacts with CBP itself, *in vitro* binding was assayed with bacterially produced p53 and GST–CBP proteins. As shown in Fig. 2a, 50% of input wild-type p53 was bound to immobilized GST–CBP3 after only a 30-min incubation. In contrast, a p53 mutant protein [p53(22,23)] (ref. 21) containing a double point mutation in the transactivation domain did not bind detectably to CBP. p53 binds tightly to CBP, with an apparent dissociation constant, K_d , of 2.5×10^{-8} M. The same analysis also revealed a weaker interaction between p53 and CBP2, consistent with the affinity-chromatography results (Fig. 1b).

To test for interaction between p53 and CBP *in vivo*, cell extracts from U2OS cells were immunoprecipitated by polyclonal antibodies (anti-CBP1 and anti-CBP3) directed against the CBP-1 and CBP-3 domains shown in Fig. 1a. p53 was detected in the immunoprecipitate with both crude anti-CBP1 and antigen affinity-purified anti-CBP1 (Fig. 2b, lanes 5 and 6), but not in either of the immunoprecipitates obtained with the preimmune sera (lane 2) or with another control antiserum (anti-SRB7) (lane 3), nor in the immunoprecipitate with anti-CBP1 from the p53-negative SaoS-2 cells (lane 7). This indicates that a specific p53–CBP complex is present in U2OS cells (Fig. 2b). Anti-CBP3 only weakly precipitates CBP–p53 immune complexes from U2OS extracts (Fig. 2b, lane 4), which may reflect an overlap between the epitope recognized by the antisera and the domain of CBP bound by p53.

The functional consequence of this interaction between p53 and CBP was investigated by cotransfection of SaoS-2 cells with the corresponding expression vectors and a reporter construct containing the *mdm-2* promoter, one of the natural targets of p53 (refs 3, 22). As shown in Fig. 3a, overexpression of CBP increased the level of activation by p53 from 5-fold up to 32-fold over the basal activity of the *mdm-2* promoter. As SaoS-2 cells do not express p53 endogenously, the synergistic effect is easily detected when as little as 2 ng of CMV–p53 plasmid is transfected into the cell (Fig. 3b). By contrast, overexpression of CBP alone has no effect on this promoter (Fig. 3b).

It has been shown that E1a downregulates p53-mediated transcriptional function on the synthetic promoter pOLXALuc²³, although the molecular mechanism of the repression is not understood. As E1a binds to CBP and inhibits CBP-dependent coactivator activity^{5–14}, we tested whether E1a-mediated repression of transactivation by p53 is dependent on the interaction with endogenous

CBP. Figure 3c confirms that overexpression of wild-type E1a can inhibit p53-mediated transactivation on the *mdm-2* reporter construct, whereas a CR1 region mutant of E1a (E1a(Δ64–68)), which is expressed at normal levels in the cell but is defective in binding to CBP/p300 (refs 10, 24), is impaired in the repression. Taken together, these data indicate that CBP functions as a coactivator to potentiate p53-dependent activation on a target gene and that E1a-mediated repression of p53 transactivation functions may be mediated, at least in part, through sequestration of endogenous CBP from p53.

According to our results (Fig. 2), p53 interacts both *in vitro* and *in vivo* with the C-terminal portion of CBP (Fig. 1a). This glutamine-rich region has a transactivation function when fused to the DNA-binding domain of Gal4 (ref. 6). Cotransfection analysis revealed that p53 increases transactivation by Gal4–CBP(1678–2441) in a dose-dependent manner, but has no effect on the transactivation function of Gal4–VP16 (Fig. 4). Although conditions in this experiment were not physiological with respect to the way in which p53 and CBP are normally anchored to the promoter, the data nevertheless indicate that tethering either CBP or p53 to the DNA is sufficient for synergism. This also suggests a role for p53 in transcriptional activation that is independent of its role in recruiting CBP to the promoter.

Finally, as the transactivation domain of p53 is important for interaction with CBP (Fig. 1c), mutations in this region should help

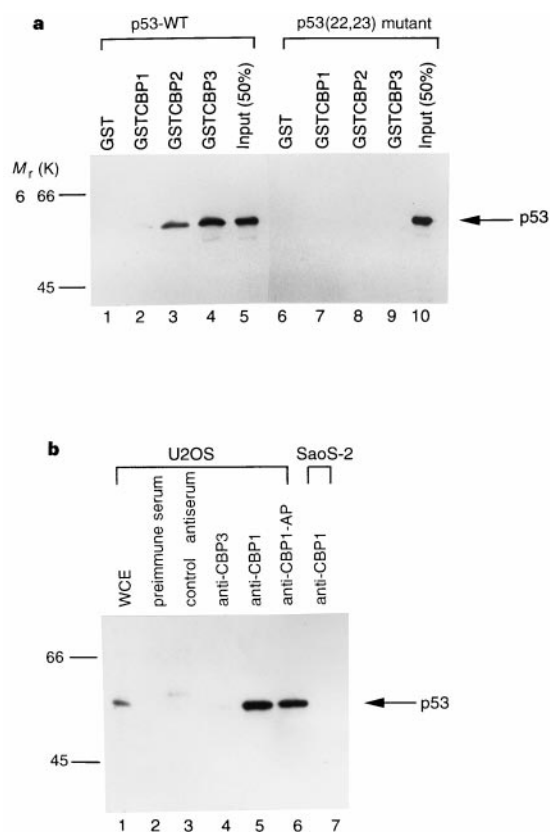


Figure 2 CBP interacts with p53 both *in vitro* and *in vivo*. **a**, Direct interactions of CBP with p53(WT) and p53(22,23) mutant proteins *in vitro*. Western blot analysis of the indicated GST pull-down complexes with anti-p53 (Ab 421). **b**, Interactions of CBP with p53 in U2OS cells. Western blot analysis of U2OS whole-cell extract (WCE) (~1% of the total protein used for immunoprecipitation) (lane 1), and the immune complexes precipitated by preimmune serum (lane 2), control antiserum (anti-SRB7) (lane 3), anti-CBP2 (lane 4), anti-CBP1 (lane 5) and anti-CBP1-AP (antigen affinity-purified) (lane 6), together with a negative control of immunoprecipitated complexes by anti-CBP1 from p53-negative cells (SaoS-2).

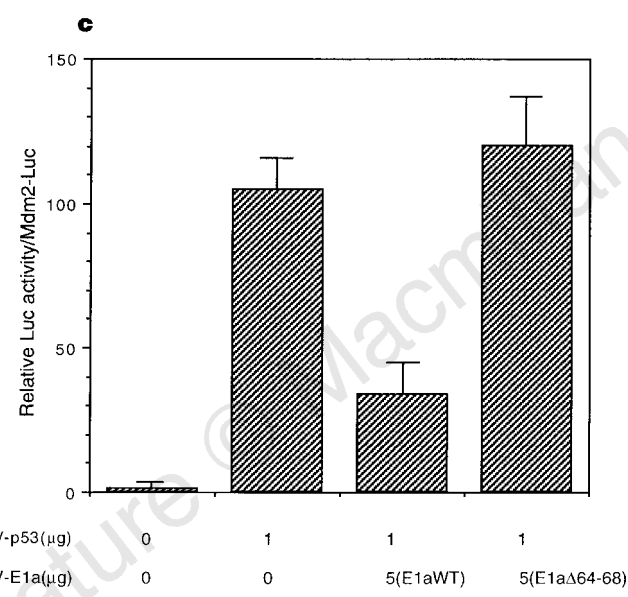
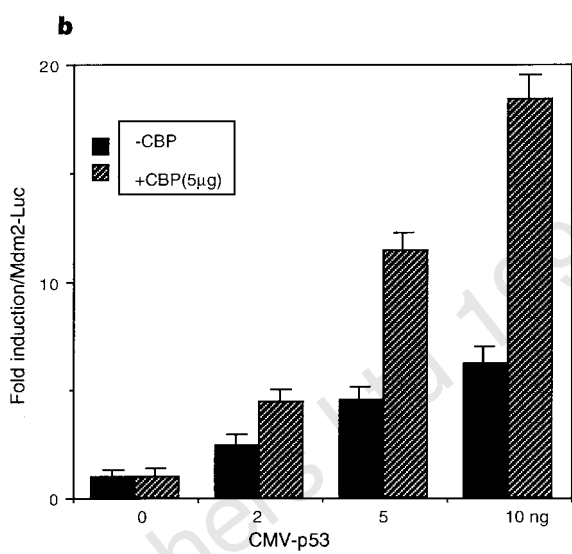
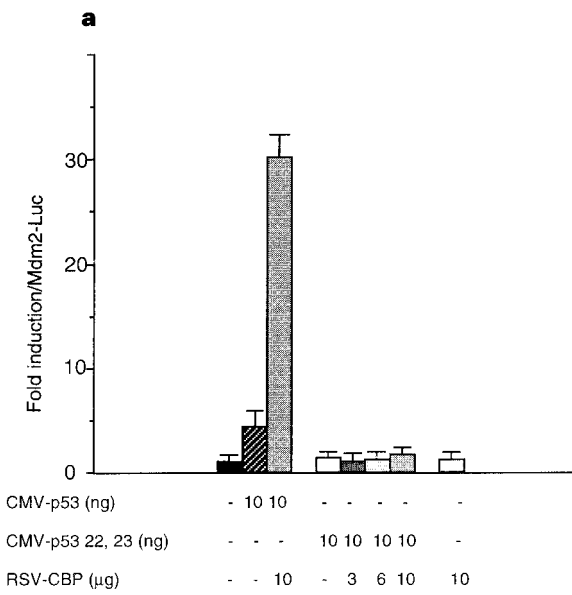


Figure 3 CBP enhances p53-mediated transactivation. **a**, CBP potentiates p53-mediated transactivation of the *mdm-2* promoter. SaoS-2 cells were transiently transfected with 10 μg Mdm-2-Luc reporter construct²², 10 ng CMV-p53(WT) or CMV-p53(22,23) (ref. 21) and 3–10 μg RSV-CBP (ref. 9), as indicated. Both CMV-p53 wild type and CMV-p53(22,23) constructs were gifts from A. Levine and are expressed at comparable levels in the cell²¹. **b**, Synergism between CBP and p53. SaoS-2 cells were transiently transfected with 10 μg Mdm-2-Luc reporter construct²², 0–10 ng CMV-p53 wild type²¹, or 5 μg RSV-CBP (ref. 9), as indicated. **c**, CBP-enhanced p53 transcriptional activity is repressed by E1a. SaoS-2 cells were cotransfected with 10 μg Mdm-2-Luc reporter construct, 1 μg CMV-p53(WT), and 5 μg 12S E1a or E1a(Δ64–68) (refs 10, 24) plasmids, as indicated.

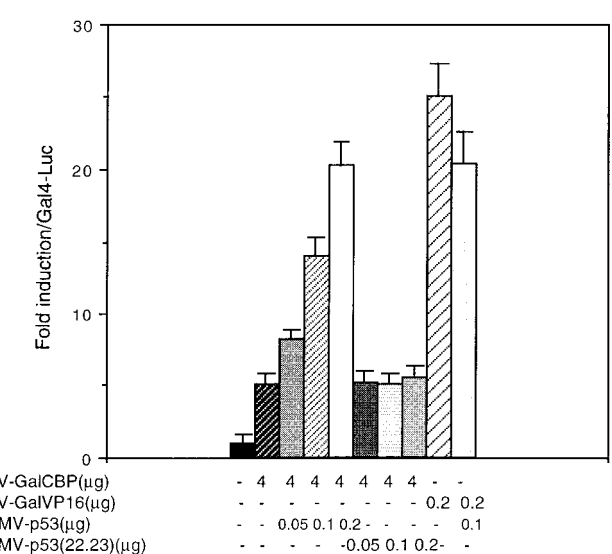


Figure 4 Synergism between Gal-CBP(1678–2441) and p53. SaoS-2 cells were transiently cotransfected with 4 μg Gal-Luc reporter construct²³, 4 μg RSV-Gal-CBP(1678–2441) (ref. 6), 50 to 200 ng CMV-p53(WT) or CMV-p53(22,23) mutant plasmids²¹ or 200 ng SV-Gal-VP16, as indicated. 3 μg CMV-LacZ was included in each case as an internal control³⁰.

to elucidate the specificity of this interaction. As predicted, a p53 mutant (p53(22,23)) that is defective in CBP binding (Fig. 2a) is severely impaired in functional synergy with CBP. Thus, as shown in Fig. 3a and Fig. 4, p53(22,23) fails to show synergy either with CBP on the *mdm-2* promoter (Fig. 3a) or with Gal4–CBP on the Gal4-target promoter (Fig. 4). Therefore, the interaction of these two proteins correlates with their functional synergy.

Our results indicate that CBP is required for optimal p53-mediated transactivation. CBP participates in preventing the G0/G1 transition during the cell cycle, by activating certain enhancers and stimulating differentiation pathways^{4–14}, whereas the tumour-suppressor gene product p53 is a key regulator of cell proliferation, differentiation and apoptosis^{1–3,25}. p53 may be one of the targets for CBP in mediating its cellular functions. As sequence-specific transcriptional activation by p53 correlates with its ability to suppress cell growth^{1–3}, cell-growth suppression by CBP may be mediated, at least in part, through synergistic activation with p53 on its target genes. CBP is also a target of adenovirus E1a (refs 5–14 and this work) and our results indicate that the binding of E1a to CBP could suppress the synergistic effect of CBP and p53 on transcriptional activation, which could also be essential for the transformation activity of this oncoprotein.

CBP/p300 is involved in transcriptional activation by several cellular factors, although the precise mechanisms by which these activators stimulate the transcriptional machinery through CBP/p300 are not yet clear. A number of activators that interact with CBP (refs 4–14, and this work) interact both physically and functionally with other coactivators: these coactivators include specific TAFs in the case of p53 (refs 17, 18) and CREB (ref. 26), members of the Src family^{12,13,27,28} and potentially other factors (reviewed in refs 12, 13) in the case of nuclear receptors, and a special set of coactivators (TRAPs) in the case of the thyroid hormone receptor²⁹. Optimal function of the activators requires both CBP/p300 and the other coactivators, suggesting that these are important but have distinct roles. This idea is supported by the discovery that both p300/CBP and an interacting factor (P/CAF) have histone acetyltransferase activity^{5,15,16}, which could modify chromatin structure. The fact that p53 exists as a tetramer *in vivo* should allow it to interact simultaneously with CBP and other coactivators (TAFs). This may be essential for optimal activation *in vivo* and also explains the synergy between CBP and p53, even when CBP is tethered to the promoter by an alternative mechanism. □

Methods

Plasmids and fusion proteins. To construct GST fusion proteins, DNA sequences corresponding to the indicated regions of CBP were amplified by PCR and subcloned into pGEX-2T (Pharmacia). GST fusion proteins were expressed in *E. coli*, extracted with buffer BC0 (20 mM Tris-HCl, pH 8.0, 0.5 mM EDTA, 20% glycerol, 1 mM DTT and 0.5 mM PMSF) containing 500 mM KCl and 1% NP-40, and purified on glutathione–Sephadex (Pharmacia). Plasmids encoding p53 wild-type or mutant proteins were created by amplification of the appropriate DNA fragments, including the Flag sequence, followed by subcloning into pET-11d (Novagene). The p53 proteins were expressed in BL21 (Lys) cells and purified on an M2 agarose affinity column (IBI).

Affinity chromatography. Affinity columns were prepared by immobilizing GST or GST–CBP proteins on glutathione–Sephadex as described¹⁹. 400-μl aliquots of HeLa nuclear extract (~4 mg protein) were loaded onto 20-μl affinity columns containing different GST fusion proteins. After washing the beads with 200 μl buffer BC0 containing 100 mM KCl and 0.1% N-P40, bound proteins were eluted with the BC0 buffer containing 1M KCl, and the total eluates and 0.5% of the input nuclear extract were subjected to western blotting. For the GST–p53 affinity columns the procedures were basically the same, except that 100-μl aliquots of HeLa nuclear extract were loaded on the affinity columns, and only half of the total proteins eluted from the columns were used for western blot analysis.

In vitro binding assays. 100 ng of p53(WT) or p53 mutant proteins were

incubated at 4°C for 30 min with 1 μg of each of the different GST fusion proteins in 500 μl total volume. This was followed by incubation with 10 μl glutathione–Sephadex beads for another 15 min. Beads were then washed six times in 1 ml BC0 buffer containing 200 mM KCl and 0.2% N-P40. Bound proteins were eluted with an equal volume of SDS loading buffer, resolved by SDS–PAGE, and analysed by western blot.

Coimmunoprecipitation assay. About 2×10^8 human osteosarcoma U2OS (or SaoS-2) cells were extracted with 5 ml lysis buffer (25 mM HEPES-KOH, pH 8.0, 150 mM KCl, 2 mM EDTA, 1 mM DTT, 1 mM PMSF, $10 \mu\text{g ml}^{-1}$ aprotinin, $10 \mu\text{g ml}^{-1}$ leupeptin, $1 \mu\text{g ml}^{-1}$ pepstatin A, 20 mM NaF, 0.1% NP-40). After centrifuging twice, the supernatants were incubated with different antisera (as indicated), crosslinked to protein A–Sephadex (Pharmacia) for 4 h at 4°C. The beads were washed six times with 1 ml lysis buffer, after which the associated proteins, including p53, were eluted with buffer BC0 containing 1 M KCl and 1% deoxycholate. Elution was for 30 min at 0°C to avoid significant elution of antibodies and associated antigens, and possible interference in the western blot result as a result of the similar size of the immunoglobulin proteins and p53. Anti-CBP1 and anti-CBP3 antisera were prepared by immunizing rabbits with the corresponding CBP-derived recombinant proteins (Fig. 1a).

Transient transfection assays. SaoS-2 cells (1×10^6) were transfected by calcium phosphate precipitation as described³⁰, and β-galactosidase and luciferase were assayed 30 h later. All transfections were done in duplicate; representative experiments depict the average of three experiments with standard deviations indicated.

Immunoblotting. Proteins were resolved on either 6% SDS–PAGE (for CBP) or 10% SDS–PAGE (for p53), followed by transfer to nitrocellulose membranes for 14 h at 4°C. After incubation with the indicated antibodies for 1 h at room temperature, blots were subsequently incubated with secondary antibodies and visualized by ECL as suggested by the manufacturer (Amersham). All procedures were basically the same as for the coimmunoprecipitation experiments, except that the blot was incubated with anti-p53 (DO-1) for 12 h at 4°C.

Received 30 October 1996; accepted 17 March 1997.

- Ko, L. J. & Prives, C. p53: puzzle and paradigm. *Genes Dev.* **10**, 1054–1072 (1996).
- El-Deiry, W. S. *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817–825 (1993).
- Wu, X., Bayle, J. H., Olson, D. & Levine, A. J. The p53–mdm-2 autoregulatory feedback loop. *Genes Dev.* **7**, 1126–1132 (1993).
- Chrivia, J. C. *et al.* Phosphorylated CREB binds specifically to the nuclear protein CBP. *Nature* **365**, 855–859 (1993).
- Yang, X. J., Ogryzko, V. V., Nishikawa, J., Howard, B. H. & Nakatani, Y. A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1a. *Nature* **382**, 319–324 (1996).
- Kwok, R. P. *et al.* Nuclear protein CBP is a coactivator for the transcription factor CREB. *Nature* **370**, 223–226 (1994).
- Arias, J. *et al.* Activation of cAMP and mitogen responsive gene relies on a common nuclear factor. *Nature* **370**, 226–229 (1994).
- Arany, Z., Newsome, D., Oldread, E., Livingston, D. M. & Eckner, R. A family of transcriptional adaptor proteins targeted by the E1a oncoprotein. *Nature* **374**, 81–84 (1995).
- Lundblad, J. R., Kwok, R. P., Lurance, M. E., Harter, M. L. & Goodman, R. H. Adenoviral E1a-associated protein p300 as a functional homologue of the transcriptional coactivator CBP. *Nature* **374**, 85–88 (1995).
- Banister, A. J. & Kouzarides, T. CBP-induced stimulation of c-Fos activity is abrogated by E1a. *EMBO J.* **14**, 4758–4762 (1995).
- Dai, P. *et al.* CBP as a transcriptional coactivator of c-Myb. *Genes Dev.* **10**, 528–540 (1996).
- Kamei, Y. *et al.* A CBP integrator complex mediates transcriptional activation and AP-1 inhibition by nuclear receptors. *Cell* **85**, 403–414 (1996).
- Chakravarti, D. *et al.* Roles of CBP/p300 in nuclear receptor signalling. *Nature* **383**, 99–103 (1996).
- Lee, J. S., See, R. H., Deng, T. & Shi, Y. Adenovirus E1a downregulates c-Jun and JunB-mediated transcription by targeting their coactivator p300. *Mol. Cell. Biol.* **16**, 4312–4326 (1996).
- Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivator p300 and CBP are histone acetyltransferase. *Cell* **87**, 953–959 (1996).
- Bannister, A. J. & Kouzarides, T. The CBP coactivator is a histone acetyltransferase. *Nature* **384**, 641–643 (1996).
- Thut, C., Chen, J. L., Klemm, R. & Tjian, R. p53 transcriptional activation mediated by coactivators TAFII40 and TAFII60. *Science* **267**, 100–104 (1995).
- Lu, H. & Levine, A. J. Human TAF31 protein is a transcriptional coactivator of the p53 protein. *Proc. Natl Acad. Sci. USA* **92**, 5154–5158 (1995).
- Xiao, H. *et al.* Binding of basal transcription factor TFIIF to the acidic activation domains of VP16 and p53. *Mol. Cell. Biol.* **14**, 7013–7024 (1994).
- Seto, E. *et al.* Wild-type p53 binds to the TATA-binding protein and represses transcription. *Proc. Natl Acad. Sci. USA* **89**, 12028–12032 (1992).
- Lin, J., Chen, J., Elenbaas, X. & Levine, A. J. Several hydrophobic amino acids in the p53 amino-terminal domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 E1b 55-kD protein. *Genes Dev.* **8**, 1235–1246 (1994).
- Haupt, Y., Rowan, S., Shaulian, E., Vowsden, K. & Oren, M. Induction of apoptosis in HeLa cells by transactivation-deficient p53. *Genes Dev.* **9**, 2170–2183 (1995).
- Stegenga, W. T. *et al.* Adenovirus E1a proteins inhibit activation of transcription by p53. *Mol. Cell. Biol.* **16**, 2101–2109 (1996).
- Wong, H. K. & Ziff, E. B. Complementary functions of E1a conserved region 1 cooperate with conserved region 3 to activate adenovirus serotype 5 early promoters. *J. Virol.* **68**, 4910–4920 (1994).
- Godley, L. A. *et al.* Wild-type p53 transgenic mice exhibit altered differentiation of the ureteric bud and possess small kidneys. *Genes Dev.* **10**, 836–850 (1996).

26. Ferreri, K., Gill, G. & Montminy, M. The cAMP-regulated transcription factor CREB interacts with a component of the TFIID complex. *Proc. Natl Acad. Sci. USA* **91**, 1210–1213 (1994).
27. Onate, S. A., Tsai, S. Y., Tsai, M.-J. & O'Malley, B. W. Sequence and characterization of a coactivator for the steroid hormone receptor superfamily. *Science* **270**, 1354–1357 (1995).
28. Voegel, J. J., Heine, M. J. S., Zechel, C., Chambon, P. & Gronemeyer, H. TIF2, a 160 kD transcriptional mediator for the ligand-dependent activation function AF-2 of nuclear receptors. *EMBO J.* **15**, 3667–3675 (1996).
29. Fondell, J. D., Ge, H. & Roeder, R. G. Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex. *Proc. Natl Acad. Sci. USA* **93**, 8329–8333 (1996).
30. Gu, W., Bhatia, K., Magrath, I. T., Dang, C. V. & Dalla-Favera, R. Binding and suppression of the Myc transcriptional activation domain by p107. *Science* **264**, 251–254 (1994).

Acknowledgements. We thank R. H. Goodman, A. Levine, M. Oren, A. G. Jochemsen, N. C. Jones, A. J. Banister and T. Kouzarides for plasmids; H. Xiao, Y. Tao, L. Wang, S. Stevens and J. D. Fondell for discussions and for critical comments on the manuscript; and Y. Nakatani for sharing unpublished observations. This work was supported by a postdoctoral fellowship from Life Science Foundation for Advanced Cancer Studies to W.G., and by grants from the NIH to R.G.R.

Correspondence and requests for materials should be addressed to R.G.R. (e-mail: roeder@rockvax.rockefeller.edu).

Binding and modulation of p53 by p300/CBP coactivators

Nancy L. Lill*, Steven R. Grossman, Doron Ginsberg*, James DeCaprio & David M. Livingston

The Dana-Farber Cancer Institute and Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA

* Present addresses: Department of Rheumatology and Immunology, Brigham and Women's Hospital, 514 Seeley G. Mudd Building, 250 Longwood Avenue, Boston, Massachusetts 02115, USA (N.L.L.); Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, 76100, Israel (D.G.).

The adenovirus E1A and SV40 large-T-antigen oncoproteins bind to members of the p300/CBP transcriptional coactivator family. Binding of p300/CBP is implicated in the transforming mechanisms of E1A and T-antigen oncoproteins. A common region of the T antigen is critical for binding both p300/CBP and the tumour suppressor p53 (ref. 1), suggesting a link between the functions of p53 and p300. Here we report that p300/CBP binds to p53 in the absence of viral oncoproteins, and that p300 and p53 colocalize within the nucleus and coexist in a stable DNA-binding complex. Consistent with its ability to bind to p300, E1A disrupted functions mediated by p53. It reduced p53-mediated activation of the *p21* and *bax* promoters, and suppressed p53-induced cell-cycle arrest and apoptosis. We conclude that members of the p300/CBP family are transcriptional adaptors for p53, modulating its checkpoint function in the G1 phase of the cell cycle and its induction of apoptosis. Disruption of p300/p53-dependent growth control may be part of the mechanism by which E1A induces cell transformation. These results help to explain how p53 mediates growth and checkpoint control, and how members of the p300/CBP family affect progression from G1 to the S phase of the cell cycle.

When stabilized by T antigen, p53 was found to bind members of the p300/CBP family (p300, CBP and p400)^{1,2}. We investigated whether non-T antigen-mediated stabilization of p53 enables p53–

p300/CBP complexes to be detected. Complex formation was assessed in ts20TG^R cells³, which are temperature sensitive in the E1 ubiquitin-activating function, and accumulate stable p53 at 39.5°C. Endogenously labelled p400 and CBP co-immunoprecipitated with p53 (Fig. 1a, compare lanes 4 and 8 with 2; also data not shown), and p53 co-immunoprecipitated with p300/CBP family members (Fig. 1a, lanes 6 and 7) at the non-permissive temperature. Unlike CBP, p300 co-immunoprecipitated with p53 at both temperatures by p300-specific immunoblotting (Fig. 1b, lanes 4 and 6), suggesting that detection of p300/p53 binding is not dependent on p53 metabolic stabilization. Taken together, the data show that stable p53–p300/CBP family-member complexes form in the absence of T antigen. A comparison of the quantities of p53-coprecipitated p300 and total p300 available (Fig. 1b) suggests that ≤1% of total cellular p300 coprecipitated with p53 under the conditions used.

E1A is known to inhibit p53 transcription-activation function⁴, and binds to a specific domain (C/H3) of p300/CBP⁵. Given that p53 also binds p300/CBP family members, we investigated whether E1A inhibits p53 transcription by binding C/H3 and downregulating p300/CBP-mediated p53 coactivation. U-2 OS cells, which synthesize wild-type p53, were transfected with a consensus p53 binding site-containing CAT reporter (PG₁₃-CAT), or its mutant counterpart (MG₁₅-CAT), which is not responsive to p53 (ref. 6). Wild-type 12S E1A specifically inhibited p53-mediated activation of PG₁₃-CAT (Fig. 2a). E1A mutant Δ2-36 failed to repress p53-dependent transcription. This mutant binds to pRB, but not to p300/CBP family members⁷. The E1A mutant, CXd1, which binds p300/CBP but not proteins of the retinoblastoma family, fully repressed promoter activity. These data indicate that one or more p300/CBP family members may coactivate p53. Mutant E1A Δ26–35 is defective in binding to p400 but not p300 (ref. 8), yet E1A Δ26–35 actively repressed p53-dependent transcription. Hence p400 is not the only p300 family member responsible for p53 coactivation.

PG₁₃-CAT and two other p53-responsive reporter plasmids, pWVP-luc, which carries the p21^{WAF1/cip1} promoter⁹, and pTM667-3, which carries the *bax* promoter¹⁰, were transfected with a p53 expression plasmid into p53-null Saos-2 cells (Fig. 2b–d). Again, E1A repression of p53-mediated transcription activation correlated with its ability to bind p300/CBP family members. Hence the same genetics of E1A repression of p53 transcription activation apply to two naturally occurring p53-activated promoters. Therefore, p300/CBP family members are implicated as modulators of p53-dependent transcription-activation function.

Overproduction of wild-type p300 override E1A-mediated repression of PG₁₃-CAT (Fig. 2e). The mutant p300 species, del33, which lacks an intact C/H3 domain and cannot bind to E1A⁵, failed to override repression. This is consistent with the view that E1A represses p53-mediated transcription activation by binding p300/CBP family member(s). The data are also consistent with a model in which the C/H3 domain of p300/CBP contributes to p53 coactivation, and E1A inactivation of this domain is relieved through sequestration of E1A by overexpressed p300. C/H3 is also the

Table 1 E1A override of p53-induced G1 cell-cycle arrest and apoptosis

Experiment code	Transfected DNA	Transfected cells (%)			
		sub-G1 (apoptotic)	G1	S	G2/M
A	pCMV	7.24	21.4	26.5	52.1
	p53	17.72	43.9	14.3	41.9
	p53 + wt E1A	7.30	28.8	17.1	54.2
	p53 + E1A Δ 2–36	14.08	41.4	19.3	39.4
B	pCMV	7.66	31.5	36.4	32.0
	p53	18.04	47.2	16.7	36.0
	p53 + wt E1A	10.80	34.5	29.3	36.2
	p53 + E1A Δ 2–36	16.10	45.3	21.5	33.3

Results are for Saos-2 cells. The G1, S and G2/M values are percentages of the non-apoptotic transfected cell population.

binding site for the histone acetylase P/CAF¹¹, suggesting that p53–p300/CBP transcription signalling might involve chromatin remodelling by P/CAF.

To determine whether a p300 family member is a component of a stable p53–DNA-binding complex, electrophoretic mobility shift assays (EMSA) were performed (Fig. 3a), using as the DNA probe a p53-binding site from the p21 promoter, together with nuclear extract from Saos-2 cells producing ectopic p53. A p53–DNA complex was observed (lane B, open arrow). Complex formation was p53 dependent (data not shown) and sequence specific (compare lanes C and D). The complex was specifically supershifted by the p53-specific antibodies pAb 421 (not shown) and pAb 1801 (compare lanes B and E, filled arrow). Anti-p300/CBP antibodies¹² TAP p300L and AC 238 eliminated the p53–DNA complex (lanes F and G). The p300 antibody RW 128 did not perturb the complex (lane H), possibly because of epitope masking. Importantly, AC238 specifically super-supershifted the pAb 1801-supershifted complex (compare lanes J and N, bracketed region). Taken together, these data suggest that nearly all of the DNA-binding p53 detected in this assay is complexed with a p300/CBP family member. E1A failed to disrupt any of these complexes (lanes P, Q and S). Therefore, disruption of the complex is probably not the basis for E1A inhibition of p53-mediated transactivation.

The existence of p53–p300/CBP-family-member complexes was further supported by the intracellular localization of p53 and p300 in cotransfected U-2 OS cells (Fig. 3b–d). Green fluorescent

protein-tagged p53 (GFP–p53) yielded a diffuse nuclear green fluorescence (Fig. 3b, left). Cotransfection of haemagglutinin-tagged wild-type p300 brought GFP–p53 into nuclear ‘dots’ (Fig. 3c, left), to which p300 colocalized (Fig. 3c, middle and right). When overproduced, p300 exists in such dots^{5,13}; p130 is excluded from them¹³, implying that they are p300 specific. When wild-type E1A was cotransfected with GFP–p53 and epitope-tagged p300, neither p300 nor p53-containing dots were detected (Fig. 3d, right and left, respectively). In contrast, cotransfection of $\Delta 2$ –36 E1A, which is unable to bind to p300, together with GFP–p53 and epitope-tagged p300, failed to disrupt either the p300 dots or their recruitment of GFP–p53 (data not shown). Similar results were obtained in Saos-2 cells (data not shown). These data reveal a strong correlation between recruitment of p53 into p300 nuclear dots and the ability of p53 specifically to transactivate certain target promoters, such as *p21* and *bax*.

Finally, we investigated whether two p53-associated biological functions, G1 arrest and induction of apoptosis, were mediated by p300. Overproduction of wild-type p53 in Saos-2 cells induces G1 arrest¹⁴. This effect was relieved by cotransfection of wild-type 12S E1A, but not by $\Delta 2$ –36 (Table 1). Similarly, wild-type E1A, but not $\Delta 2$ –36 E1A, suppressed p53-induced apoptosis at early times after transfection (Table 1). Therefore, the p300/CBP contribution to p53 transcription function may underlie the ability of p53 to arrest cells in G1 and to induce apoptosis. It has been suggested that p53 can induce apoptosis through two mechanisms, one of which is dependent

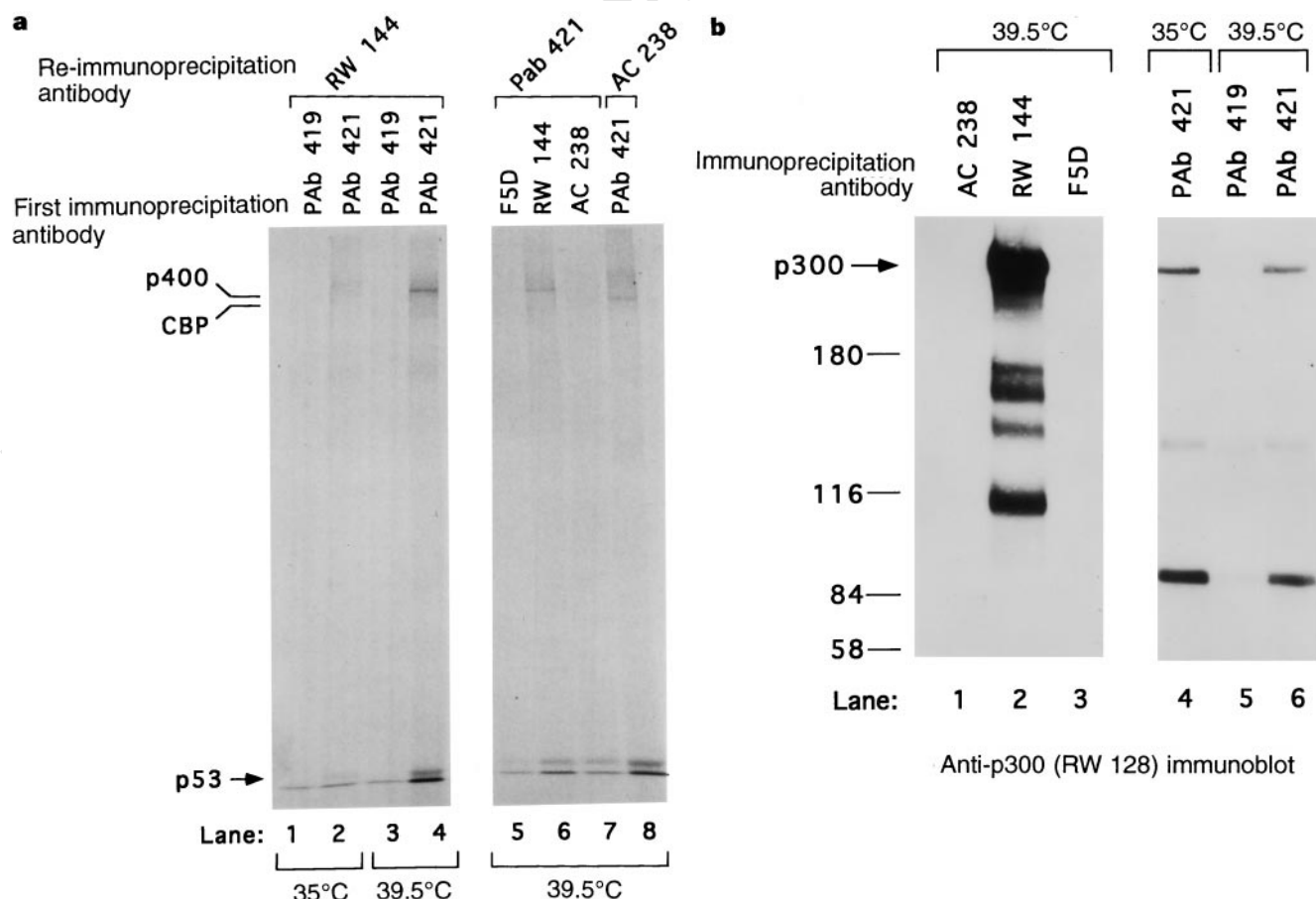


Figure 1 p53–p300 family member complexes form without viral oncoproteins. **a**, Immunoprecipitation/re-immunoprecipitation of ³⁵S-labelled proteins; ts20TG^R protein amounts were 2 (39.5°C) or 3.8 mg (35°C). Reactions using 35°C extracts contained half the p53 present in those using 39.5°C extracts (not shown). **b**, Immunoprecipitation/RW 128 immunoblotting. Immunoprecipitates were

prepared from unlabelled ts20TG^R extract proteins (35°C, 25 mg, lane 4; 39.5°C, 15 mg, lanes 5 and 6). RW 128 immunoblotted p300 (lane 2) but not CBP (lane 1). Lanes 1–3 show immunoprecipitates from 150 µg of 39.5°C extract protein. The protein of ~90K in lanes 4 and 6 is unidentified, and may be a p300 degradation product.

on p53 transcription activation^{15,16}. The induction of early apoptosis (<48 h) by p53 is transactivation-domain dependent, whereas late apoptosis (>72 h) is not¹⁶. E1A-mediated suppression of apoptosis in Saos-2 cells was apparent at 48 h after transfection and was barely observable 72 h after transfection (data not shown). Therefore, p53/p300 complex formation may significantly influence transcription-dependent early apoptosis.

Our results strongly suggest that the ability of p53 to mediate important functions, such as its DNA-damage/G1-checkpoint-activation function, is p300/CBP dependent. p300/CBP function has been linked independently to the control of both G1 exit and genome stability^{17,18}. These two regulatory activities probably depend on an effective checkpoint-maintenance function in which p53–p300/CBP complexes seem to be important. It will be interesting to determine whether p300/CBP directly mediate(s) p53 transcription signalling to the core transcription apparatus and more specifically, whether this involves the known interactions^{19,20} of p300/CBP and p53 with TBP, TAFs or P/CAF¹¹.

Note added in proof: While this manuscript was under review, it was reported that E1A inhibits p53-mediated p21 transactivation and cell-cycle arrest through its p300/CBP-binding domain³⁰. □

Methods

Plasmids. Reporters PG₁₃-CAT⁶, MG₁₅-CAT⁶ and pWVP-luc⁹, and wild-type p53 expression vector pC53-SN-3 (ref. 21) were provided by B. Vogelstein. The *bax* reporter, pTM667-3 (ref. 10), was provided by J. Reed. The pCMV12S²² and

pE1A.CXdl vectors⁷ were provided by E. White. pCMV12S encodes wild-type 12S E1A. The E1A CXdl mutant lacks amino acids 121–150. E1A Δ2–36 and Δ26–35 coding sequences were derived from plasmids provided by E. Moran⁷ and S. Bayley⁸, respectively. The pCD-GFPp53 vector encoding GFP–p53 was generated by insertion of a GFP cassette²³ (provided by P. Silver) into expression vector pCDNA3-p53 (provided by W. G. Kaelin Jr and C. Jost). The CD20 expression vector, pCMVCD20 (ref. 24), was provided by L. Zhu. pBSK has been described (Stratagene, pBluescript SK).

Cells. The ts20TG^R cells, a gift of H. Ozer, have been described³. Saos-2 and U-2 OS human osteosarcoma cell lines were obtained from the American Type Culture Collection.

Antibodies. The anti-p300/CBP ascites used in this study were described previously^{5,12}. TAPp300L was provided by M. Ewen. AC 238 and RW 128 ascites recognize both p300 and CBP in EMSA assays¹². For immunoprecipitations, hybridoma supernatant AC 238 (IgG1) is CBP specific, whereas RW 144 (IgG1) is p300/p400 specific. Hybridoma supernatant RW128 is p300 specific in immunoblotting. The antibodies PAb 421 (IgG2a, anti-p53) and pAb419 (control IgG2a, anti-SV40 T antigen) were provided by E. Harlow. Anti-myogenin antibody F5D (control IgG1) was provided by W. Wright. Anti-CD20 antibody was obtained from Dako (M0774). Antibody 12CA5 (BAbCo) recognizes the haemagglutinin epitope. Anti-p53 antibody pAb 1801 was obtained from Santa Cruz.

Immunoprecipitation/re-immunoprecipitation. Experiments were performed as described¹, using cells incubated at the indicated temperatures for 16 h.

Transfection. Transfections were performed using the calcium phosphate

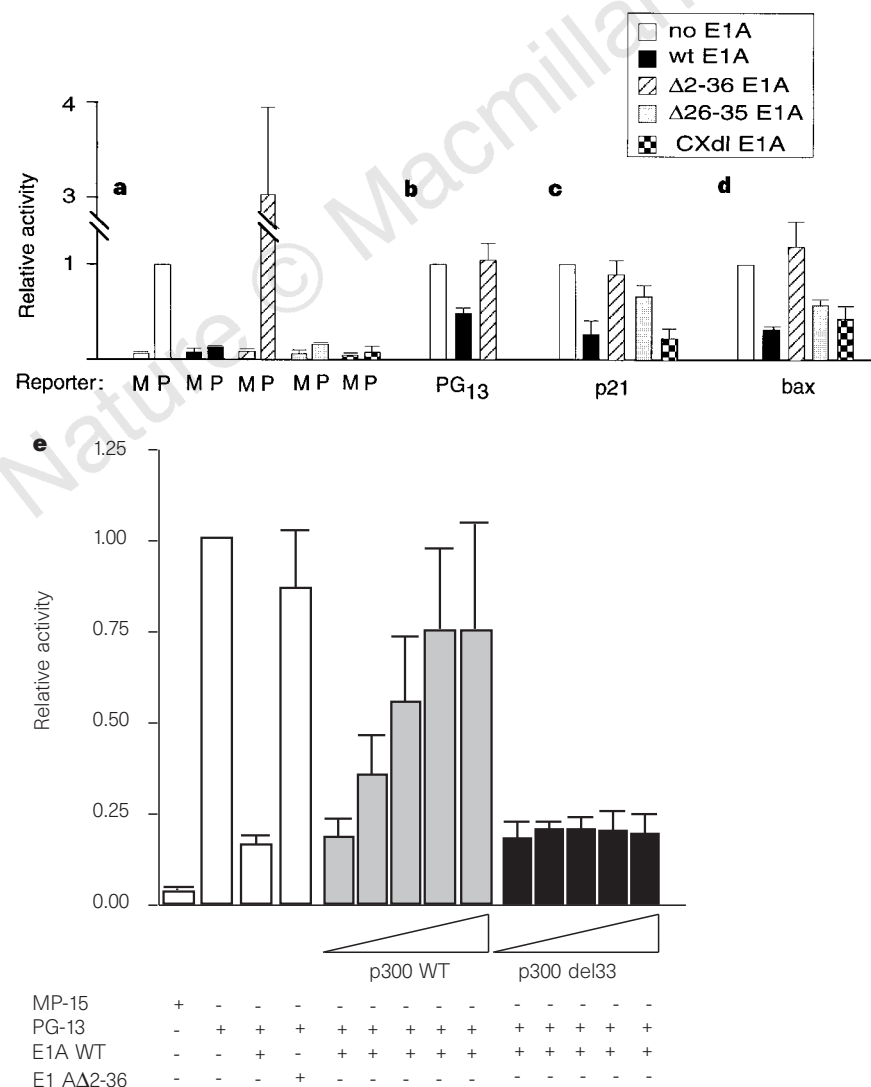
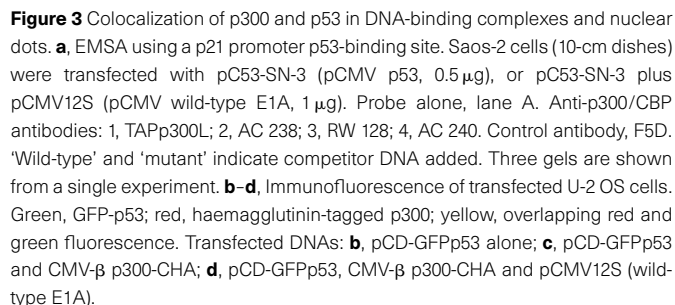


Figure 2 Genetics of E1A-mediated inhibition of p53-responsive reporters. **a**, U-2 OS cells received 2 μg of reporter (M, MG₁₅; P, PG₁₃), 0.5 μg of E1A expression vector, or equimolar pCMV vector. DNA (7.5 μg) was added to 60-mm dishes. **b–d**, Saos-2 cells (10-cm dishes) were transfected with 5 μg reporter, 1 μg E1A expression vector, and pC53-SN-3 (pCMV p53; 2 μg for **b**; 0.5 μg for **c** and **d**) or equimolar pCMV vector; 18.5 μg of DNA was added per dish. Reporter plasmids were: **b**, PG₁₃-CAT; **c**, pWVP-luc; **d**, pTM667-3. Results are averages of three experiments for **a–c**, and two experiments for **d**. For **b** and **d**, relative activity in the absence of p53 was less than 0.01. For **c**, the highest activity in the absence of p53 was 0.13. **e**, U-2 OS cells (10-cm dishes) were transfected with 5 μg reporter and 0.1 μg E1A expression vector; 17.1 μg DNA was added per dish. Increasing amounts (0.1, 2, 6, 10 and 12 μg) of p300 expression vector were added as shown. Ectopic proteins (wild-type p300 (WT) and p300 del33) were comparably expressed (not shown). Results are averages of three experiments, each with duplicate samples. Error bars, s.d.



Cell-cycle and apoptosis analysis. Soas-2 cells were transfected with (μg DNA per 10-cm dish): 0.5 pC53-SN-3, 1.0 pCMV12S or Δ 2-36, 1.0 pCMVCD20. Total DNA added per dish was 18.5 μg . Cells were immunostained for cell-surface CD20 and propidium iodide-stained for DNA²⁹. For FACS analysis, DNA content of CD20⁺ cells was evaluated. Cells with sub-G1 content were designated the apoptotic population.

1. Lill, N. L. *et al.* p300 family members associate with the carboxyl terminus of simian virus 40 large tumor antigen. *J. Virol.* **71**, 129–137 (1997).
2. Eckner, R. *et al.* Association of p300 and CBP with simian virus 40 large T antigen. *Mol. Cell. Biol.* **16**, 3454–3464 (1996).
3. Chowdary, D. R. *et al.* Accumulation of p53 in a mutant cell line defective in the ubiquitin pathway. *Mol. Cell. Biol.* **14**, 1997–2003 (1994).
4. Steegenga, W. T. *et al.* Adenovirus E1A proteins inhibit activation of transcription by p53. *Mol. Cell. Biol.* **16**, 2101–2109 (1996).
5. Eckner, R. *et al.* Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev.* **8**, 869–884 (1994).
6. Kern, S. E. *et al.* Oncogenic forms of p53 inhibit p53-regulated gene expression. *Science* **256**, 827–830 (1992).
7. Stein, R. W. *et al.* Analysis of E1A mediated growth regulation functions binding of the 300-kilodalton cellular product correlates with E1A enhancer repression function and DNA synthesis-inducing activity. *J. Virol.* **64**, 4421–4427 (1990).
8. Howe, J. A. & Bayley, S. T. Effects of Ad5 E1A mutant viruses on the cell cycle in relation to the binding of cellular proteins including the retinoblastoma protein and cyclin A. *Virology* **186**, 15–24 (1992).
9. El-Deiry, W. S. *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817–825 (1993).
10. Miyashita, T. & Reed, J. C. Tumor suppressor p53 is a direct transcriptional activator of the human *bax* gene. *Cell* **80**, 293–299 (1995).
11. Yang, X. J. *et al.* A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A. *Nature* **382**, 319–324 (1996).
12. Eckner, R. *et al.* Interaction and functional collaboration of p300/CBP and bHLH proteins in muscle and B-cell differentiation. *Genes Dev.* **10**, 2478–2490 (1996).
13. Yao, T.-P. *et al.* The nuclear hormone receptor coactivator SRC-1 is a specific target of p300. *Proc. Natl Acad. Sci. USA* **93**, 10626–10631 (1996).
14. Diller, L. *et al.* p53 functions as a cell cycle control protein in osteosarcomas. *Mol. Cell. Biol.* **10**, 5772–5781 (1990).
15. Haupt, Y. *et al.* Induction of apoptosis in HeLa cells by transactivation-deficient p53. *Genes Dev.* **9**, 2170–2183 (1995).
16. Chen, X. *et al.* p53 levels, functional domains, and DNA damage determine the extent of the apoptotic response of tumor cells. *Genes Dev.* **10**, 2438–2451 (1996).
17. Bayley, S. T. & Mymryk, J. S. Adenovirus E1A proteins and transformation. *Int. J. Oncol.* **5**, 425–444 (1994).
18. Caporossi, D. & Bachetti, S. Definition of the adenovirus type 5 functions involved in the induction of chromosomal aberrations in human cells. *J. Gen. Virol.* **71**, 801–808 (1990).
19. Abraham, S. E. *et al.* p300, and p300-associated proteins, are components of TATA-binding protein (TBP) complexes. *Oncogene* **8**, 1639–1647 (1993).
20. Ko, L. J. & Privés, C. p53: puzzle and paradigm. *Genes Dev.* **10**, 1054–1072 (1996).
21. Baker, S. J. *et al.* Suppression of human colorectal carcinoma cell growth by wild-type p53. *Science* **249**, 912–915 (1990).

22. White, E. *et al.* Adenovirus E1B 19-kilodalton protein overcomes the cytotoxicity of E1A proteins. *J. Virol.* **65**, 2968–2978 (1991).
23. Kahana, J. & Silver, P. in *Current Protocols in Molecular Biology* (eds Ausubel, F. M. *et al.*) 9.7.22–9.7.28 (Greene and Wiley Interscience, New York, 1996).
24. Zhu, L. *et al.* Inhibition of cell proliferation by p107, a relative of the retinoblastoma protein. *Genes Dev.* **7**, 1111–1125 (1993).
25. Ausubel, F. M. *et al.* *Current Protocols in Molecular Biology* (Greene and Wiley Interscience, New York, 1996).
26. Krek, W., Livingston, D. M. & Shirodkar, S. Binding to DNA and the retinoblastoma gene product promoted by complex formation of different E2F family members. *Science* **262**, 1557–1560 (1993).
27. Chittenden, T., Livingston, D. M. & DeCaprio, J. A. Cell cycle analysis of E2F in primary human T cells reveals novel E2F complexes and biochemically distinct forms of free E2F. *Mol. Cell. Biol.* **13**, 3975–3983 (1993).
28. Friedlander, P. *et al.* A mutant p53 that discriminates between p53-responsive genes cannot induce apoptosis. *Mol. Cell. Biol.* **16**, 4961–4971 (1996).
29. Qin, X.-Q. *et al.* The transcription factor E2F-1 is a downstream target of RB action. *Mol. Cell. Biol.* **15**, 742–755 (1995).
30. Somasundaram, K. & El-Deiry, W. S. Inhibition of p53-mediated transactivation and cell cycle arrest by E1A through its p300/CBP-interacting region. *Oncogene* **14**, 1047–1057 (1997).

Acknowledgements. We thank everyone who provided reagents used in this study; J. Kahana, P. Silver, C. Jost and W. G. Kaelin Jr for providing unpublished reagents; P. Adams, R. Eckner, M. Ewen, O. Iliopoulos, R. Scully, T.-P. Yao and members of the Division of Neoplastic Disease Mechanisms for discussions; M. Modabber for graphics; and M. Simone for flow cytometry. This work was supported by grants from the American Cancer Society (N.L.L.), the Howard Hughes Medical Institute (S.R.G.), The Cancer Research Fund of the Damon Runyon–Walter Winchell Foundation (D.G.), the Dana-Farber/Sandoz Drug Discovery Program, and the National Cancer Institute (J. DeC. and D.M.L.).

Correspondence and requests for materials should be addressed to D.M.L. (e-mail: david_livingston@dfci.harvard.edu).

Structure of isopenicillin N synthase complexed with substrate and the mechanism of penicillin formation

Peter L. Roach[†], Ian J. Clifton[†], Charles M. H. Hensgens^{*}, Norio Shibata^{*}, Christopher J. Schofield^{*}, Janos Hajdu[†] & Jack E. Baldwin^{*}

^{*} The Dyson Perrins Laboratory and the Oxford Centre of Molecular Sciences, University of Oxford, South Parks Road, Oxford OX1 3QY, UK

[†] Department of Biochemistry, Biomedical Centre, Uppsala University, Box 576, S-751 23 Uppsala, Sweden

The biosynthesis of penicillin and cephalosporin antibiotics in microorganisms requires the formation of the bicyclic nucleus of penicillin¹. Isopenicillin N synthase (IPNS), a non-haem iron-dependent oxidase, catalyses the reaction of a tripeptide, δ -(L- α -aminoadipoyl)-L-cysteinyl-D-valine (ACV), and dioxygen to form isopenicillin N and two water molecules². Mechanistic studies suggest the reaction is initiated by ligation of the substrate thiolate to the iron centre, and proceeds through an enzyme-bound monocyclic intermediate^{3,4} (Fig. 1). Here we report the crystal structure of the IPNS complexed to ferrous iron and ACV, determined to 1.3 Å resolution. Based on the structure, we propose a mechanism for penicillin formation that involves ligation of ACV to the iron centre, creating a vacant iron coordination site into which dioxygen can bind. Subsequently, iron-dioxygen and iron-oxo species remove the requisite hydrogens from ACV without the direct assistance of protein residues (Fig. 2). The crystal structure of the complex with the dioxygen analogue, NO and ACV bound to the active-site iron supports this hypothesis.

Spectroscopic studies of IPNS in the resting state have suggested the presence of two or three histidines, an aspartate and possibly two water molecules as metal ligands^{5,6}. The crystal structure⁷ of *Aspergillus nidulans* IPNS complexed with divalent manganese (substituting for iron) revealed a metal ion octahedrally coordinated by His 214, Asp 216, His 270, Gln 330 and two water molecules (Fig. 3a). Under aerobic conditions, crystals of IPNS bound with both iron and ACV could not be obtained owing to instability

and turnover problems. Crystals of IPNS complexed to ferrous iron and ACV were therefore grown under anaerobic conditions⁸.

The Fe(II):ACV:IPNS structure has one protein molecule with ferrous ion and ACV bound at the active site in the asymmetric unit (Fig. 3b). Substrate binding does not distort the 'jelly-roll' core of the enzyme, in which the iron and ACV are enclosed. The side chain of Gln 330, which coordinates the metal in the absence of substrate, is replaced by the ACV thiolate. In the Fe(II):ACV:IPNS complex, the seven carboxy-terminal residues adopt a conformation that extends the final helix (α -10) relative to the Mn:IPNS structure and encloses the substrate in the active site.

The ACV is anchored within the active site by ligation of its thiolate to the iron centre and through its two carboxylate groups (Fig. 4a). One of the two water molecules ligating the metal ion in the Mn:IPNS complex is displaced, changing the metal coordination geometry from octahedral to square pyramidal (Fig. 2: 4 $\rightarrow$ 5). In the substrate complex, three of the five coordination sites are filled with protein ligands: His 214, His 270 and Asp 216 (ref. 9). The remaining two sites are occupied by a water molecule (at position 398) and the ACV thiolate. The valine isopropyl group is held in van der Waals contact with the iron by interactions with Leu 231, Val 272, Pro 283 and Leu 223. The presence of the substrate-derived valine methyl group in the coordination site *trans* to Asp 216 prevents a water molecule from binding. Remarkably, the valine β C–H bond, which is cleaved during the formation of the thiazolidine ring, is directed away from the iron centre. The pentacoordinate and high-spin nature of the Fe(II):ACV:IPNS complex, and the displacement of one water molecule on ACV binding, were previously inferred from spectroscopic studies^{5,6,10}.

The aminoadipoyl residue of ACV lies in an extended conformation, as predicted from analogue studies^{1,3,11}, and its carboxylate makes a salt bridge with Arg 87, replacing that formed between the C-terminal carboxylate (Thr 331) and Arg 87 in the Mn:IPNS structure¹².

The carboxylate of the valine residue of ACV is prevented from coordinating to the iron centre by a hydrogen-bonding network. Before ACV binds, the side chain of Arg 279 points out of the active site towards the exterior of the enzyme (Fig. 3a), but in the Fe(II):ACV:IPNS structure (Figs 3b and 4a), Arg 279 is directed into the active-site cavity by forming a hydrogen bond to the valine carboxylate of ACV through a bridging water molecule (labelled as

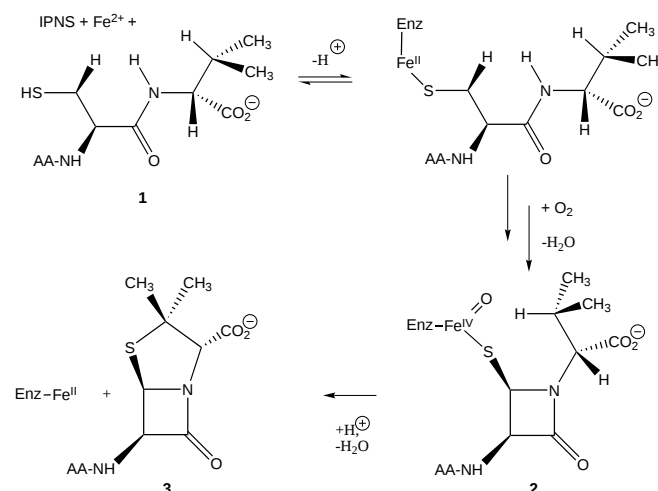


Figure 1 The IPNS reaction pathway through the proposed enzyme (enz)-bound monocyclic intermediate (2). AA, L- δ -(α -aminoadipoyl).

Table 1 Data collection and statistics

	Fe:(II):ACV:IPNS		Fe:NO:ACV:IPNS	
X-ray source	BM14, ESRF, Grenoble		ID2, ESRF, Grenoble	
R_{cryst} (%) [*]	18.2 (8–1.30 Å)		20.66 (25–1.45 Å)	
R_{free} (%) [†]	19.8		21.2	
Unit Cell	$a = 46.8, b = 71.5, c = 101.0 \text{ Å}$		$a = 47.26, b = 73.08, c = 101.94 \text{ Å}$	
Resolution shell	1.35–1.30 Å	25.0–1.3 Å	1.45–1.49 Å	22.0–1.45 Å
Measurements	25,182 [‡]	267,069	10,327	180,036
Average $I/\sigma I$	4.1	25.9	2.3	21.2
Unique reflections	7,714	80,233	3,298	50,924
Completeness (%)	93.2	95.4	85.8	92.0
$I/\sigma I > 3$ data only (%)	52.2	79.2	30.3	69.4
R_{merge} (%) [§]	31.8	5.5	51.9	8.2

^{*} $R_{\text{cryst}} = \sum ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum |F_{\text{obs}}| \times 100$ for the resolution range shown. Figures are for an isotropic, individual B -factor model.

[†] R_{free} based on 4% of the total reflections.

[‡]Estimated value.

[§] $R_{\text{merge}} = \sum_j \sum_h |I_{hj} - \langle I_h \rangle| / \sum_j \sum_h \langle I_h \rangle \times 100$, where I_{hj} is the intensity of an individual measurement and $\langle I_h \rangle$ is the mean intensity of that reflection.

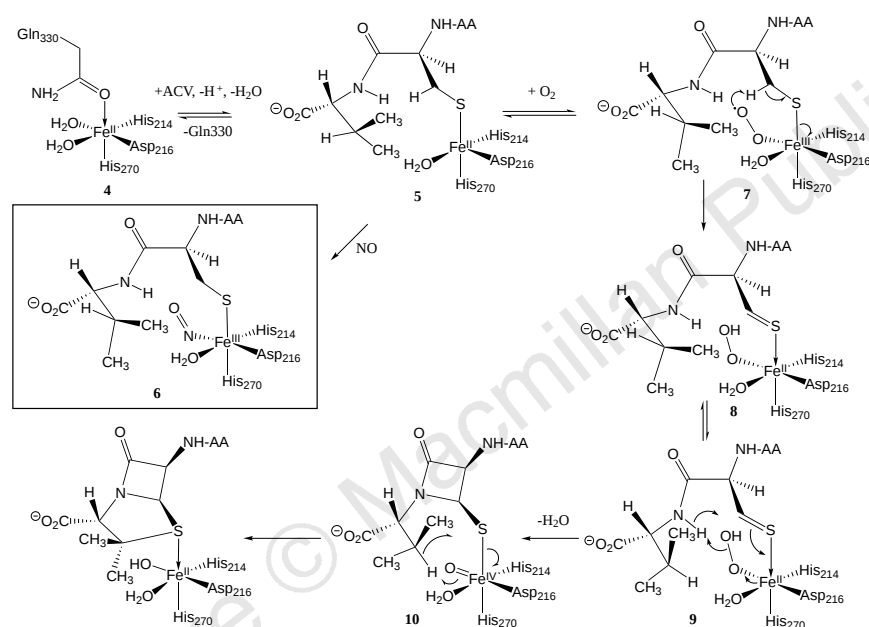


Figure 2 Mechanism for isopenicillin *N* formation and the formation of the Fe:NO:ACV:IPNS complex.

352 in Fig. 4a). The ACV valine carboxylate also forms hydrogen bonds with the side-chain hydroxyls of Ser 281, Tyr 189 and a water molecule (labelled as 567 in Fig. 4a).

Electron paramagnetic resonance (EPR) studies⁵ using the dioxygen analogue nitric oxide have indicated that the NO binds to the Fe(II):ACV:IPNS complex to yield an octahedral complex. When exposed to NO, the crystals of Fe(II):ACV:IPNS also reacted to form the ternary Fe:NO:ACV:IPNS complex, which has a visible spectrum similar to that observed in solution (λ_{max} for crystals, 507, 715 nm; for solution, 508, 720 nm)^{10,13}. The crystal structure of the Fe:NO:ACV:IPNS complex reveals that NO binds to the iron centre *trans* to Asp 216 (Fig. 4b), suggesting that this is the likely oxygen-binding site. To accommodate the NO molecule, a rotation of $\sim 30^\circ$ occurs around the substrate valine C α –C β bond, increasing the distance between the iron and the ACV-derived valine methyl carbon from 3.90 to 4.81 Å. The NO is bound in a nonlinear orientation (Fe–N–O angle of 120°) and the NO oxygen atom is equidistant (3.3 Å) from both the valine nitrogen and cysteinyl β -carbon, which must each donate a hydrogen atom to dioxygen to close the β -lactam ring.

Ligation of ACV to the iron presumably results in a reduction of the Fe(II)/Fe(III) redox potential, allowing dioxygen to bind, thereby initiating the reaction cycle (Fig. 2). The simplest way to integrate the spectroscopic and structural results is by dioxygen binding in the

site *trans* to Asp 216 (5 $\rightarrow$ 7). With a nonlinear, haemoglobin-like binding of dioxygen, the superoxide in 7 is juxtaposed to the *pro*-3-S hydrogen of the ACV cysteinyl residue, and could remove this hydrogen atom (7 $\rightarrow$ 8)¹⁴. Subsequent cleavage of the hydroperoxide with concomitant deprotonation of the amide N–H allows simultaneous β -lactam closure and ferryl formation (9 $\rightarrow$ 10). Rotation of the isopropyl group (8 $\rightarrow$ 9) must occur before or at the same time as β -lactam formation to relieve its steric interactions with the sulphur ligand. This rotation directs the valine β -hydrogen towards the ferryl-oxo species with which it reacts^{3,4,15}.

The ACV is bound rigidly (ACV average B -factor = 9.3 Å^2 ; IPNS main-chain average B -factor = 9.9 Å^2) and is held in a conformation in which the cysteinyl carbonyl group eclipses the adjacent C α –N bond. A torsion angle of -55° is defined by the valine N and by the cysteinyl C-2, C-2 and C-3 of ACV. The tripeptide conformation is therefore approaching that required for β -lactam formation. Examination of the active site fails to reveal any residues capable of providing general acid–base catalysis. Thus the dioxygen has three main roles: as an electron acceptor during β -lactam ring formation; as a source of the ferryl species for the energetically demanding second-ring closure; and to remove specific hydrogen atoms from the substrate directly.

Comparison of the Mn:IPNS and Fe(II):ACV:IPNS structures shows that, when ACV binds to IPNS, Phe 211 undergoes a 120°

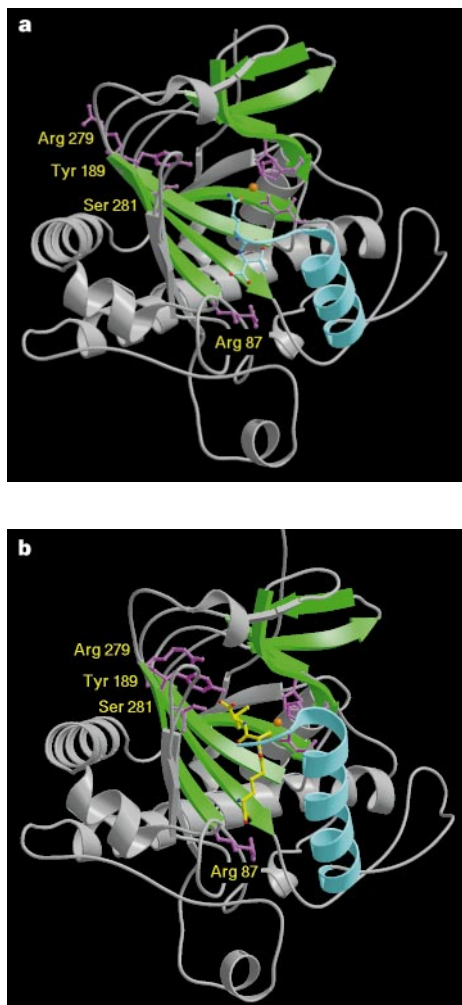


Figure 3 Comparison of the structures of Mn:IPNS (**a**) and Fe(II):ACV:IPNS (**b**) from the same orientation. The jelly-roll motif is in green, the C-terminal region (residues 313–331) cyan, the active-site metal ion (manganese in **a**, iron in **b**) orange, the key substrate-binding residues (His 214, His 270, Asp 216, Arg 87, Arg 279, Tyr 189 and Ser 281) magenta, and the ACV yellow.

rotation about its C α –C β bond, moving the phenyl ring to form part of the active-site cavity (Fig. 3a). In this position, it may help to isolate the proposed hydroperoxide intermediate from potential hydrogen-bonding partners, enhancing the basicity of the hydroperoxide-hydroxide and facilitating deprotonation of the ACV valine N–H. The movement of Phe 211 opens a channel to the protein surface that is bordered by the side chains of Thr 331, Val 100 and Tyr 189 and contains three water molecules (labelled in Fig. 4a as 567, 543 and 605). The water molecule produced when the β -lactam is formed (9 $\rightarrow$ 10) may leave the active site through this channel.

Peroxide and ferryl intermediates have also been proposed in cytochrome P-450 oxygenase and oxidase reactions¹⁶ in which a cysteinyl thiolate may also be the immediate source of electrons for ferryl formation¹⁷. However, the different geometrical relationships between bound dioxygen and sulphur in the haem and non-haem systems means that they have different reaction pathways. In IPNS, the *cis* sulphur–dioxygen relationship at the iron centre allows the

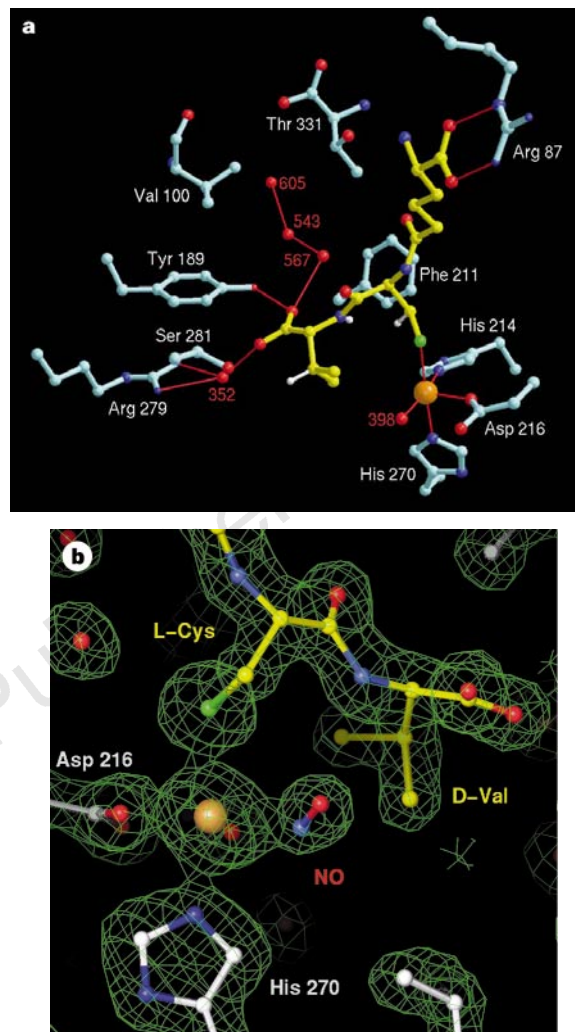


Figure 4 Structure of the active-site region. **a**, The three hydrogen atoms removed to form isopenicillin *N* are shown in white. The iron-sulphur bond length of 2.41 Å observed in this crystal structure is slightly longer than that observed either in solution-phase EXAFS studies⁶ of *Cephalosporium acremonium* IPNS 2.32 ± 0.02 Å or in EXAFS studies⁸ of the crystalline *A. nidulans* IPNS 2.34 ± 0.03 Å. **b**, Binding of NO to the iron centre. The $2F_o - F_c$ electron density map is contoured at 1.13σ .

involvement of the ACV-derived cysteinyl moiety in the redox chemistry leading to the β -lactam; in the P-450 enzymes, the *trans* relationship isolates the thiolate from dioxygen so that removal of the 3-cysteinyl hydrogen is not possible. In both cases, however, substrate binding displaces water to create a pentacoordinate iron species before the binding of oxygen. Consequently, although both porphyrin- and non-porphyrin-based catalysts may have evolved to carry out similar oxidative reactions (for example, hydroxylations and desaturations), there may be other reactions (for example, ACV to isopenicillin *N*) which can only be catalysed by non-haem-based catalysts.

Many (but not all) 2-oxoglutarate-dependent oxygenases and 1-aminocyclopropane-1-carboxylic-acid oxidase (ACCO) show significant sequence similarity to IPNS^{7,14,18}. In addition to conservation of the metal-binding ligands corresponding to His 214, Asp 216 and His 270 in IPNS and the predicted tertiary structure (the jelly-roll motif), other residues are conserved across the IPNS subfamily, including Arg 279 and Ser 281, which bind to the ACV valine

carboxylate (Fig. 4a). This carboxylate-binding motif is probably involved in binding the C-5 carboxylate of 2-oxoglutarate, the cosubstrate for all known members of the subfamily except IPNS and ACCO, allowing the C-1 carboxylate of 2-oxoglutarate to coordinate to the iron atom¹⁹.

Sequences reported for Fe(II)-dependent extradiol dioxygenases^{20–22} display significant sequence similarity to each other, but not to those of the IPNS subfamily. However, crystallographic studies have shown that the coordination chemistry and proposed mechanism of the extradiol dioxygenases share common features with IPNS. The structure of the Fe(III) form of 2,3-dihydroxybiphenyl dioxygenase with a substrate bound reveals two imidazole (His 209 and His 145) and one carboxylate (Glu 260) metal ligands, with the proposed oxygen-binding site *trans* to the glutamate and the substrate ligated directly to the iron²². Thus the IPNS subfamily and the extradiol dioxygenases may have convergently evolved similar solutions to the mechanistic problems posed by using dioxygen as an oxidant. □

Methods

Data collection. All data were collected at 100 K using 0.997-Å radiation and a 30-cm MAR research detector (Table 1).

Structure determination of Fe(II):ACV:IPNS. The Fe(II):ACV:IPNS crystals belong to the space group $P2_12_12_1$. Data were processed with the DENZO and SCALEPACK programs²³, and initial phases were calculated by molecular replacement using the program AMoRe²⁴. Electron density maps were interpreted using the program O²⁵. In 14 cycles of refinement, using the programs XPLOR²⁶, PROLSQ²⁷ and SHELXL93 (ref. 28), 328 residues (4–331) were fitted to the electron density. In the final cycle, the positions of all the atoms in the asymmetric unit, including 322 water molecules, a sulphate ion, the ferrous ion and ACV were refined using SHELXL93 which gave a crystallographic *R*-factor of 13.8% (calculated using anisotropic temperature factors).

Structure determination of Fe:NO:ACV:IPNS. Crystals were prepared under anaerobic conditions by transferring a single coverslip with a drop containing Fe(II):ACV:IPNS crystals to a fresh Linbro crystallization tray, injecting NO gas (1 ml) under the coverslip, and rapidly resealing the well. The crystals reacted by diffusion over 1 h and developed an orange–pink colour. Data from these crystals were processed with MOSFILM²⁹ and the CCP4 suite of programs³⁰. An initial structure was obtained by rigid body refinement of the Fe(II):ACV:IPNS main-chain residues into the new unit cell. In 9 cycles of refinement using REFMAC³⁰, 327 residues (5–331) were fitted to the electron density. In the final cycle, 116 water molecules, iron, NO and ACV were refined. Electron density for the NO was clearly visible throughout the refinement, and the NO was modelled in after two cycles. For both structures, the iron–ligand bond lengths were unrestrained throughout the refinement, and there were no Ramachandran outliers.

Received 29 January; accepted 15 April 1997.

1. Baldwin, J. E. & Abraham, E. Biosynthesis of penicillins and cephalosporins. *Nat. Prod. Rep.* **5**, 129–145 (1988).
2. Pang, C. P. *et al.* Purification of isopenicillin N synthetase. *Biochem. J.* **222**, 789–795 (1984).
3. Baldwin, J. E. & Schofield, C. J. in *The Chemistry of β -lactams* (ed. Page, M. I.) 1–78 (Blackie, London, 1992).
4. Que, L. & Ho, R. Y. N. Dioxygen activation by enzymes with mononuclear non-haem iron active sites. *Chem. Rev.* **96**, 2607–2624 (1996).
5. Orville, A. M. *et al.* Thiolate ligation of the active site iron(II) of isopenicillin N synthase derives from substrate rather than endogenous cysteine: spectroscopic studies of site-specific Cys $\rightarrow$ Ser mutated enzymes. *Biochemistry* **31**, 4602–4612 (1992).
6. Randall, C. R. *et al.* X-ray absorption studies of the ferrous active site of isopenicillin N synthase and related model complexes. *Biochemistry* **32**, 6664–6673 (1993).
7. Roach, P. L. *et al.* Crystal structure of isopenicillin N synthase is the first from a new structural family of enzymes. *Nature* **375**, 700–704 (1995).

8. Roach, P. L. *et al.* Anaerobic crystallisation of an isopenicillin N synthase:Fe(II):substrate complex demonstrated by X-ray studies. *Eur. J. Biochem.* **242**, 736–740 (1996).
9. Borovok, I., Landman, O., Kreisberg-Zakarin, R., Aharonowitz, Y. & Cohen, G. Ferrous active site of isopenicillin N synthase: genetic and sequence analysis of endogenous ligands. *Biochemistry* **35**, 1981–1987 (1996).
10. Chen, V. J. *et al.* Spectroscopic studies of isopenicillin N synthase. A mononuclear nonhaem Fe²⁺ oxidase with metal coordination sites for small molecules and substrate. *J. Biol. Chem.* **264**, 21677–21681 (1989).
11. Baldwin, J. E. *et al.* Penicillin biosynthesis: active site mapping with aminoadipoylcysteinylvaline variants. *J. Chem. Soc. Chem. Commun.* 1225–1227 (1984).
12. Rowe, C. J. thesis, Oxford Univ. (1995).
13. Hadfield, A. & Hajdu, J. A fast and portable microspectrophotometer for protein crystallography. *J. Appl. Crystallogr.* **26**, 839–842 (1993).
14. Cooper, R. D. G. The enzymes involved in biosynthesis of penicillin and cephalosporin: Their structure and function. *Bioorg. Med. Chem.* **1**, 1–17 (1993).
15. Baldwin, J. E. *et al.* Evidence for an insertion-homolysis mechanism for carbon–sulfur bond formation in penicillin biosynthesis; 2. Incubation and interpretation. *Tetrahedron* **52**, 2537–2556 (1996).
16. Groves, J. T. & Han, Y. Z. in *Cytochrome P-450: Structure, Chemistry and Biochemistry* 2nd edn (ed. Ortiz de Montellano, P. R.) 3–49 (Plenum, New York, 1995).
17. Baldwin, J. E., Morris, G. M. & Richards, W. G. Electron transport in cytochromes P-450 by covalent switching. *Proc. R. Soc. Lond. B* **245**, 43–52 (1991).
18. Prescott, A. G. A dilemma of dioxygenases (or where biochemistry and molecular biology fail to meet). *J. Exp. Bot.* **44**, 849–861 (1993).
19. Hanauke-Abel, H. M. & Guzzler, V. A stereochemical concept for the catalytic mechanism of prolylhydroxylase. *J. Theor. Biol.* **94**, 421–455 (1982).
20. Shu, L. *et al.* X-ray absorption spectroscopic studies of the Fe(II) active site of catechol 2,3-dioxygenase. Implications for the extradiol cleavage mechanism. *Biochemistry* **34**, 6649–6659 (1995).
21. Han, S., Eltis, L. D., Timmis, K. N., Muchmore, S. W. & Bolin, J. T. Crystal structure of the biphenyl-cleaving extradiol dioxygenase from a PCB-degrading pseudomonad. *Science* **270**, 976–980 (1995).
22. Senda, T. *et al.* Three-dimensional structures of free form and two substrate complexes of an extradiol ring-cleavage type dioxygenase, the BphC enzyme from *Pseudomonas* sp. strain KKS102. *J. Mol. Biol.* **255**, 735–752 (1996).
23. Otwinowski, Z. in *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. W. & Bailey, S.) 55–62 (Daresbury Laboratory, Warrington, UK, 1993).
24. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
25. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
26. Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallographic *R* factor refinement by molecular dynamics. *Science* **235**, 458–460 (1987).
27. Konner, J. H. & Hendrickson, W. A. A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr. A* **36**, 344–350 (1980).
28. Sheldrick, G. M. *SHELXL93, Program for Crystal Structure Refinement* (Univ. Gottingen, Germany, 1993).
29. Leslie, A. G. W. *Mosflm* (MRC Laboratory of Molecular Biology, Cambridge, 1996).
30. CCP4 The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

Acknowledgements. We thank K. Harlos, E. Garman, R. Bryan, I. Andersson, R. M. Adlington, R. C. Wilmouth, V. Fülöp, J. P. N. Pitt, A. Howe, S. Lee, J. W. Keeping, B. Rasmussen and A. Thompson for help and discussions. Financial support was provided by the MRC, BBSRC, EPSRC and Zeneca through a DTI link scheme.

Correspondence and requests for materials should be addressed to J.E.B. or J.H. The crystallographic coordinates have been deposited in the Brookhaven Protein Data Bank (accession nos 1IPS, 2IPS and 3IPS) and will be released one year after publication.

erratum

Photonic crystals: putting a new twist on light

J. D. Joannopoulos, Pierre R. Villeneuve & Shanhui Fan

Nature **386**, 143–149 (1997)

Figure 6 in this Review was shown in the wrong orientation. The figure as printed should be rotated 90° anticlockwise. □